# Moral Stability and Liberal Justification:

## An Examination of the Notion of Stability in Rawls's Theory

A thesis submitted for the degree of Doctor of Philosophy

by

**Po Chung CHOW**

Department of Government
London School of Economics and Political Science
University of London

June 2006





ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

## ABSTRACT

This thesis offers a comprehensive examination of the notion of moral stability in Rawls's political philosophy. I argue that the problem of stability is essentially concerned with the motivational priority of a sense of justice. A conception of justice is justified if and only if it can effectively motivate rational agents to act as justice requires. It is a constitutive condition of justifiability rather than a practical matter of feasibility. I vindicate my claim through a philosophical interpretation of Rawls's account of practical reason for action. I then contend that stability plays an essential role in Rawls's two stage justification. At the first stage, taking place in the original position, stability is one of the main grounds for Rawls's principles. Nevertheless, I argue that the motive for contractors to adopt the maximin rule stems from moral considerations rather than an egoistic rational choice. At the second stage, the question of how to reconcile justice and goodness arises. We need to consider whether the regulative desire to act justly is congruent with a person's own good. This concern leads to Rawls's congruence argument through a Kantian interpretation of human nature. I suggest that this interpretation has turned Rawls into a liberal perfectionist within a classical teleological framework – a position inconsistent with Rawls's desire-based conception of prudential rationality. It is this internal inconsistency which makes the congruence argument fundamentally flawed. I then turn to examine political liberalism and point out that the idea of an overlapping consensus fails to justify the priority of political values over non-political ones. Finally, I propose an idea of potential congruence to support justice as fairness as a stable conception of justice. I conclude that this is the right direction to resolve the problem of stability and justification.

## AUTHOR'S DECLARATION

I declare that the work presented in this thesis is mine alone.

Po Chung CHOW

## ACKNOWLEDGEMENTS

It has taken me eight years to finish this thesis. This intellectual journey, though unexpectedly long, has been a wonderful one. During these years, I have learnt a lot from my teachers and friends to whom I am deeply indebted.

The first person I would like to thank is John Rawls—the object of my research. Although I have never met Rawls, I have learnt how to do political philosophy and why it matters through his works. He has shaped my life and made doing political philosophy my ground project.

My greatest debt is to my supervisor, Prof. John Charvet. I first met Prof. Charvet in the summer of 1996. It was a sunny afternoon. I told him I wanted to do a Ph.D. with him. After a brief talk, he agreed to supervise me. I became his student in 1998. Since then, Prof. Charvet has been a guide in my philosophical enquiry. We have had many discussions on political philosophy in his office, coffee shops around LSE, and most importantly, in his inspiring home seminars. As a supervisor, Prof. Charvet is always open-minded, supportive and insightful. I am grateful for his careful examination on my work, his encouragement to my ideas, and his generosity as a teacher and a person. I own a great debt to my second supervisor Prof. Paul Kelly. Since I arrived at LSE, Prof. Kelly has taken good care of me. He has given me insightful and wise comments on my works, and provided me good advice on the direction of my research.

I thank my Ph.D. examiners Prof. Jonathan Wolff and Prof. Matt Matravers for their incisive comments on my thesis.

I would like to extend my heart-felt gratitude to my teacher Prof. Shih Yuan-Kang at the Chinese University of Hong Kong. Prof. Shih gave me my first taste of political philosophy in 1993 when I took his course *Liberals and Communitarians*. I first heard about Rawls from him. Over the years, we have had

numerous philosophical dialogues. He has read through my whole thesis and given me a great deal of penetrating comments. Without his enlightenment and support, I would not have stepped into the world of political philosophy to start with. I am also grateful to my teachers at the University of York, Susan Mendus, Peter Nicholson and Matt Matravers, who had given me very good philosophical training.

I treasure highly the mutual support between my friends in the Government Department and me. They include Alessandra Sarquis, Jessie Wei, Jeong Won Park, Philip Cook, Michael Bacon, Martin McIvor, Chico Gaetani, Kato Koichi, Manjeet Ramgotra, Rotem Gonen, and Hans Kribbe. As time goes by, those good old days we spent together have become my most beautiful memories. I also thank Yat-tung Chan, Wai-sang Tang, Siu-fu Tang, and Wei Wang, with whom I have continuous philosophical dialogues over the years. I am indebted to Laura Chu who has helped me proofread the whole thesis.

Some chapters have been presented in different occasions. I thank Richard Arneson, Joseph Chan, Cecile Fabre, Thomas Pogge, Joseph Raz, and Ser-min Shei for their helpful comments. I am also indebted to Thomas Scanlon for his stimulating discussion on Rawls when he visited the Chinese University of Hong Kong in 2005.

Last but not least, I thank the Overseas Research Students Awards, LSE Research Studentships, The Great Britain-China Educational Trust Studentships, and Royal Institute of Philosophy Bursary for their financial support.

This thesis is dedicated to my wife, Maggie Wan and my parents for their unfailing support and love.

# CONTENTS

# INTRODUCTION

The problem of stability is fundamental to John Rawls's political philosophy. It is the main theme of Part III of *A Theory of Justice*.[1] There he proposes an argument for congruence to resolve the problem. Rawls once said that this argument is one of the most original contributions he makes in that book.[2] Surprisingly, it attracts little attention.[3] As Freeman remarks, of all the voluminous commentary on this work, "virtually nothing has been written on the central feature of that argument [stability] on the 'congruence of the right and the good'."[4] Critics must have overlooked the significance of this problem.

The publication of Rawls's second book *Political Liberalism* has changed the situation.[5] In an important passage explaining his philosophical turn, Rawls says:

> But to understand the nature and extent of the differences, one must see them as arising from trying to resolve a serious problem internal to justice as fairness, namely from the fact that the account of *stability* in Part III of *Theory* is not consistent with the view as a whole. I believe *all differences are consequences of removing that inconsistency*. Otherwise these

---

[1] In his introduction to Part III, Rawls reminds us that "sometimes in this part the overall direction of the exposition may seem less clear, and the transition from one topic to another more abrupt. It might help to keep in mind that *the central aim is to prepare the way to settle the questions of stability and congruence*, and to account for the values of society and the good of justice." *A Theory of Justice* (Oxford: Oxford University Press, 1972), p.395; Revised Edition, 1999, p.347, my emphasis. (Referred to hereafter as TJ, with page references given parenthetically in the text, in which the first and second represent the page number of first and revised edition respectively.)

[2] Samuel Freeman, "Congruence and the Good of Justice" in *The Cambridge Companion to Rawls* ed. Samuel Freeman (Cambridge: Cambridge University Press, 2003), p.308.

[3] According to Freeman, Rawls "is puzzled why it did not attract more comment." "Congruence and the Good of Justice," p.308.

[4] Samuel Freeman, "Political Liberalism and the Possibility of a Just Democratic Constitution" *Chicago-Kent Law Review* 69, (1994), p.623.

[5] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1996), paper edition (Referred to hereafter as *PL*, with page references given parenthetically in the text).

> lectures take the structure and content of *Theory* to remain substantially the same. (PL:xvii-xviii)

Rawls ascribes the changes in his view to the problem of stability. He goes on to say that "the problem of stability has played very little role in the history of moral philosophy, so it may seem odd that an inconsistency of this kind should turn out to force such extensive revisions. Yet the problem of stability is fundamental to political philosophy and an inconsistency there is bound to require basic readjustments." (PL:xix) To critics' surprise, the problem of stability comes to the forefront of justice as fairness and becomes the cornerstone of Rawls's later development of political liberalism. So much so that if we want to understand Rawls's change, we must first understand the nature of stability, its proper role in the justificatory structure of justice as fairness, and why the later Rawls is unsatisfied with his first attempt to solve the problem in *A Theory of Justice*. I aim to probe these questions in this thesis.

A brief literature review may illuminate the significance of this project. There are different kinds of explanation for Rawls's change to political liberalism. An influential and widely-held account regards it as a response to the critique of communitarianism.[6] Few critics are convinced by Rawls's own account that his changes result from an internal inconsistency between the congruence argument

---

[6] For example, Stephen Mulhall and Adam Swift, *Liberals and Communitarians* (Oxford: Blackwell, 1992); Chandran Kukathas and Philip Pettit, *Rawls: A Theory of Justice and its Critics* (Cambridge: Polity Press, 1990); Shlomo Avineri and Avner de-Shalit ed., *Communitaranism and Individualism* (New York: Oxford University Press, 1992). Communitarian thinkers include Michael Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982); Alasdair MacIntyre, *After Virtue* (London: Duckworth, 1981); Charles Taylor, *Philosophy and the Human Sciences* vol. 2 (Cambridge: Cambridge University Press, 1985); Michael Walzer, *Spheres of Justice* (Oxford: Blackwell, 1983).

and the fact of reasonable pluralism.[7] Moreover, even those who are sympathetic to Rawls's liberal position have strong doubts about the importance of stability in general and the desirability of introducing it into justice as fairness in particular. Their reservations are based on a conventional interpretation of the problem of stability. According to this view, stability is concerned with social order and the feasibility of a conception of justice. As long as the effective means of maintaining an enduring and peaceful social cooperation can be found, the political conception is considered stable. This is a purely practical matter. Rawls takes stability seriously because he has a deep concern about how unity will be viable in a society profoundly divided by reasonable though incompatible religious, philosophical, and moral doctrines. Reasonable pluralism is a permanent fact of modern democratic society. We must take this fact into account in constructing political principles. For "the aims of political philosophy depend on the society it addresses. In a constitutional democracy one of its most important aims is presenting a political conception of justice that can not only provide a shared pubic basis for the justification of political and social institutions but also helps ensure stability from one generation to the next."[8] Rawls seems to have abandoned his universalist ambition to justify a conception of justice from the perspective that regards "the human situation not only from all social but also from all temporal points of view." (TJ:587/514 rev.) The concern for social order in an era of reasonable pluralism urges Rawls to take a practical turn to search for

---

[7] To my knowledge, the first literature to show sympathy with this account is Freeman, "Political Liberalism and the Possibility of a Just Democratic Constitution," pp.919-668. Barry also provides an intensive analysis from this perspective though he does not agree with Rawls's change. "John Rawls and the Search for Stability" *Ethics* 105, (1995), pp.874-915.

[8] Rawls, "The Idea of an Overlapping Consensus," in *Collected Papers* ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.421.

an overlapping consensus on a freestanding political conception of justice. Rawls thus defines the central problem of political liberalism as "how is it possible that deeply opposed though reasonable comprehensive doctrines may live together and all affirm the political conception of a constitutional regime?" (PL:xx) This interpretation seems to give a coherent and self-sufficient account of the causal relations between stability and the need for political liberalism.

Suppose this conventional interpretation is correct, several challenges ensue. If stability is viewed as a primary concern of political justification, justice will be substantially identified with a stable social order. Whether a conception of justice is justified depends on whether it can reduce conflict and achieve a state of peaceful and lasting cooperation. Feasibility sets a constraint on desirability at the outset. A Hobbesian pragmatism thus underlies Rawls's political liberalism.[9] If so, Rawls must pay a heavy moral cost. For by itself social stability does not define a moral point of view. It is entirely possible that a stable social order may be built on unjust social institutions. Coercive force, the threat of penalties, and internalization of the dominant class's ideology through education and brainwashing may well bring people to comply with the political rules as effectively as a sense of justice.

However, the point is not whether Rawls's proposal is practicable or not. It is rather that moral justification should not take social stability into consideration in the first place. What political philosophy should do is to look for the most desirable conception of justice justified by independent moral reasons. Whether

---

[9] Jean Hampton, "Should Political Liberalism Be Done without Metaphysics?" *Ethics* 99 (July 1989), pp.791-814.

the conception derived is stable is a separate issue that should only be taken up in the application level.

Therefore, Rawls's viewing stability as the first subject of political justification will obscure the moral character of justice as fairness and risk conceding desirability to feasibility. Worse still, it commits a category mistake. For justice and stability belongs to different categories. The latter is not a property of the former. [10] As a consequence, Rawls's whole project of political liberalism is grounded on a mistake. And given that stability is irrelevant to justification, there is no need for Rawls to shift to political liberalism even if an inconsistency is found in the congruence argument of *A Theory of Justice*. He can simply neglect it.[11]

I do not agree with this conventional interpretation. I intend to refute it and vindicate the moral importance of stability in justice as fairness. I want to argue that stability is a necessary condition of justification. Thus my thesis is entitled "moral stability and liberal justification." The thesis will focus on four related questions. (1), what does the notion of stability mean in Rawls's context? (2), why is it essential to moral justification? (3), to what extent does Rawls provide a satisfactory solution to the problem? (4), if not, what will be the alternative? The thesis will answer these questions in order. As my discussion proceeds, a different picture of stability will gradually emerge.

The first three chapters are intended to answer the first and second questions

---

[10] G. A. Cohen, "Rescuing Justice from Constructivism," typescript.

[11] This seems to be Barry's position. "John Rawls and the Search for Stability," p.883.

while the remaining three chapters are designed to deal with the third and fourth ones. In Chapter One, I draw a distinction between moral and social stability and show that Rawls's notion of stability is concerned with the motivational priority of a sense of justice. It claims that a conception of justice is justified if and only if it can effectively motivate rational agents to act as justice requires. It seeks to establish the overridingness of moral reasons. Moral stability is therefore conceptually independent of social stability. It has its own moral agenda. Once this is clear, most criticism originating from the conventional interpretation can be dismissed. Chapter Two turns to explore the importance of moral stability. I contend that Rawls has adopted a conception of prudential rationality and reason/motive internalism, which leads him to hold that the motivational priority of justice can only be made sense of by showing its regulative status in an agent's rational plan of life. Chapter Three sets out to investigate the place of stability in justice as fairness. Against Rawls and his critics, I make three major claims. First, stability is one of the main grounds for the principles of justice in the first stage. Second, the real force moving the parties in the original position to adopt the maximin rule actually results from moral considerations. Finally, the second stage is indispensable to the justification of justice as fairness. Its function is, however, not to confirm the feasibility of principles of justice, but to justify the overridingness of moral motivation.

Once the meaning of moral stability and its proper place have been settled, I proceed to examine Rawls's substantive arguments for stability. Chapter Four will assess the congruence argument. I shall first explicate the main ideas of congruence and then show that its main ground lies in a Kantian interpretation of justice as fairness. I argue that this interpretation has turned Rawls into a liberal

perfectionist within a classical teleological framework – a position inconsistent with Rawls's desire-based conception of prudential rationality. For this conception cannot warrant that rational persons would necessarily accept a Kantian interpretation of human nature. I contend that it is this internal inconsistency which makes the congruence argument fundamentally flawed. Chapter Five will turn to examine whether the idea of an overlapping consensus can provide a better alternative to resolve the problem of stability. My discussion focuses on a specific question: will a political conception of justice provide sufficient reason for a rational agent to accept the priority of political values? By examining Rawls's three arguments, including the greatness of political values, two model cases, and the idea of burdens of judgment, I argue that Rawls fails to vindicate his claim. This failure leads to the last chapter in which I propose an idea of potential congruence to support moral stability. I claim that it is rational for a person to give precedence to morality over narrow self-interest because leading a just life itself can be presented as a higher-order regulative good under favourable conditions. I first discuss two pre-conditions for potential congruence. They are the unity of practical reasoning and the pervasiveness of moral feelings. After that, I continue to argue that there are good reasons for an agent to give priority to Rawls's two principles of justice. Potential congruence is achievable in a well-ordered society. The claim of moral stability can then be vindicated.

# CHAPTER 1

# MORAL AND SOCIAL STABILITY

The problem of stability occupies a central place in Rawls's theory. It sets a normative constraint on the justifiability of a conception of justice. Rawls believes that a theory that fails to be sufficiently stable is morally unjustified. To assess Rawls's claim, the first thing we need to know is what stability means in Rawls's context, and why it is so fundamental to political philosophy. This chapter aims to explore these two questions.

This chapter consists of four sections. The first section presents Rawls's characterisation of stability and its role in the justificatory structure of justice as fairness. I call Rawls's conception of stability *moral stability*. It is concerned with the motivational priority of the sense of justice. The second section presents a predominant interpretation of stability held by most critics. This view holds that the problem of stability is primarily derived from a Hobbesian concern for social order. I dub this conception *social stability* and call the predominant view the *conventional interpretation*. According to this interpretation, the purpose of moral stability is to serve social stability. The former only has instrumental value in terms of its contribution to the latter. I shall explain why this conventional view looks plausible for many critics, and then refute it in the third and fourth sections. Section 3 shows that moral stability is not necessarily the most effective means to realize social stability. Section 4 takes up a more serious and difficult question about whether the later Rawls undergoes a fundamental change from a Kantian to a Hobbesian position by taking social stability as the first subject of justice under the challenge of reasonable pluralism. My answer is No. I argue that if we follow

the conventional interpretation, Rawls's account of stability would be normatively undesirable and conceptually absurd. If the conventional interpretation is sound, Rawls's claim that stability is fundamental to political justification would fail. Worse still, his turn to political liberalism resulting from an inconsistent account of stability would make his whole project vulnerable to further charges. So, in order to make sense of the moral significance of stability, I conclude that we should give up the conventional interpretation. In particular, we should disconnect moral stability from social stability, and search for a more coherent and attractive account of stability to make sense of its justificatory role in justice as fairness.

## 1 The Meaning of Moral Stability

According to Rawls, the problem of stability is concerned with whether a conception of justice can generate a sufficient sense of justice to win the willing compliance of citizens. The sense of justice is an effective settled desire to apply and to act from the principles of justice. A conception of justice is stable if and only if the sense of justice is shown to have regulative priority over other desires that would otherwise lead people to act unjustly. As Rawls puts it, "to insure stability men must have a sense of justice or a concern for those who would be disadvantaged by their defection, preferably both. When these sentiments are sufficiently strong to overrule the temptations to violate the rules, just schemes are stable." (TJ:497/435 rev.) When a conception is stable, the scheme of social cooperation will tend to endure over time. Even if deviations or infractions occur, citizens' settled desire to act justly will come into play to prevent further disorder, and thus restoring the just arrangement. This definition of stability remains intact throughout Rawls's writings. In *A Theory of Justice*, he states:

> One conception of justice is more stable than another if the sense of justice that it tends to generate is stronger and more likely to override disruptive inclinations and if the institutions it allows foster weaker impulses and temptations to act unjustly. The stability of a conception of justice depends upon a balance of motives: the sense of justice that it cultivates and the aims that it encourages must normally win out against propensities toward injustice. (TJ:454/398 rev.)

Rawls maintains this definition in *Political Liberalism*. What has changed is his new proposal to secure stability by reformulating justice as fairness as a political conception and grounding it on an overlapping consensus among reasonable comprehensive doctrines. But the realization of stability still depends upon:

> Citizens' sense of justice, given their traits of character and interests as formed by living under a just basic structure, is strong enough to resist the normal tendencies to injustice. Citizens act willingly so as to give one another justice over time. Stability is secured by sufficient motivation of the appropriate kind acquired under just institutions. (PL:142-43)

Several distinctive features of stability are noteworthy. First, Rawls is concerned with a special kind of stability, to wit, the stabilising force must spring from an agent's effective sense of justice. A conception of justice would not be stable if people were coerced, tricked, or pressured into acquiescing in the dictates of principles of justice. It is neither a consequence of the effective use of coercive state power nor a result of modus vivendi. "Citizens act willingly so as to give one another justice over time."(PL:142) I will call this conception *moral stability*. The first question of stability is to explain how citizens can acquire effective sense of justice to comply with political principles. This requires an account of moral

psychology. Briefly speaking, Rawls believes that rational agents beyond a certain age and possessing the requisite intellectual capacity will naturally develop a sense of justice under normal social circumstances. He even claims that "one who lacks a sense of justice lacks certain fundamental attitudes and capacities included under the notion of humanity." (TJ:488/428 rev.) Chapter 8 of *A Theory of Justice* is devoted to exploring how and under what conditions this moral sentiment is acquired. The argument draws heavily on moral psychology concerning the stages of moral learning from early age to maturity. Rawls believes that a normal agent in a just society will undergo three stages with respect to three psychological laws (the morality of authority, the morality of association, and the morality of principle) and gradually develop an effective sense of justice.

The second feature of stability is that the effectiveness of moral motivation is tied to the reasonableness of a conception of justice. In normal circumstances our reasons for action will give rise to a corresponding motive to act. The desire to act justly largely depends upon the justifying reasons for the conception of justice. While most people over a certain age develop a capacity for a sense of justice, to what extent it is regulative varies with the requirements of different theories of justice. As Nagel remarks, "the motives are not independent of political and ethical theory. Ethical argument reveals possibilities of moral motivation that cannot be understood without it, and in political theory these possibilities are elaborated through institutions to which people are able to adhere partly because of their moral attractiveness."[1] For example, the same person may find that the dictates of utilitarianism are much more demanding than those of Rawlsian

---

[1] Thomas Nagel, *Equality and Partiality* (Oxford: Oxford University Press, 1991), p.27

liberalism because the latter can better protect their rights and basic liberties while respecting the distinctness of individuals. Other things being equal, the agent can develop a stronger desire to comply with justice as fairness than with utilitarianism. As Rawls puts it, "the account of moral development is tied throughout to the conception of justice which is to be learned, and therefore presupposes the plausibility if not the correctness of this theory." (TJ:461/404 rev.) Moral stability is inseparable from the substantive content of justice. This explains why Rawls describes stability primarily as an attribute of a conception of justice rather than that of a political system.

It should, however, be noted that even if citizens have a capacity for a sense of justice, it does not mean that they will give priority to the sense of justice over other interests without qualification of any kind. Undoubtedly, justice makes claims on us. Political principles set limits to our pursuit of the good. These limits may not always be in harmony with our interest. When they conflict, a theory of justice commands the priority of moral considerations over other interests. The right is, in this sense, prior to the good. This priority of the right should be understood as a structural requirement. It does not tell us whether rational people actually have sufficient motive to act as justice requires in case of conflict. The overridingness of moral motivation is not warranted by definition. It must be supported by substantive reason. For the sense of justice is just one desire among many in an agent's motivational system. A conception of justice needs to explain how the demands of justice are derived from, compatible with, or at least not in deep conflict with people's reasonable conceptions of the good. As Freeman notes, merely showing that citizens have a sense of justice is not enough to ensure that they will consistently act justly, and that a just society will be stable. Moral

stability hinges on a satisfactory answer to the following question: "Why should they care about it sufficiently so that they have reason to subordinate pursuit of their ends to requirements of justice?"[2]

Thus the challenge of stability is to justify how a conception of justice can win the whole-hearted allegiance of citizens who have their distinct conceptions of the good. It acknowledges a tension between justice and self-interest, and then seeks to justify the priority of the former over the latter. If there is no conflict between two perspectives, or the consideration of justice is logically prior to that of the good, stability would not become an issue at all. Therefore, "the stability of a conception depends upon a balance of motives: the sense of justice that it cultivates and the aims that it encourages must normally win out against propensities toward injustice." (TJ:454/398 rev.). The overridingness of moral motivation is a necessary and sufficient condition of stability. This is the third feature of Rawls's conception of stability.

The last feature is that stability is essential to Rawls's idea of a well-ordered society. A well-ordered society is a highly idealised concept that provides a useful frame of reference to compare different conceptions of justice. Rawls assumes that a reasonable theory of justice must accord with this ideal society. A well-ordered society consists of three conditions. First, everyone accepts, and knows that everyone else accepts, the very same principles of justice; second, its basic structure generally satisfies and is publicly known to satisfy these principles; and finally its members have a strong and normally effective desire to act as the

---

[2] Samuel Freeman, "Congruence and the Good of Justice," in *The Cambridge Companion to Rawls*, ed. Samuel Freeman (Cambridge: Cambridge University Press, 2003), p.280.

principles of justice require. (TJ: 453-54/397-98 rev.; PL:35). In such a society, principles of justice and its grounds are publicly known to, and accepted by free and equal citizens. Moreover, citizens have strong inclinations to act in accordance with the publicly recognized principles. They have no intention either to violate or to renegotiate the terms of social cooperation, given their present and prospective social position. We can note that the third condition refers to moral stability. A well-ordered society must be morally stable.

Since a conception of justice must meet the requirements of a well-ordered society, it follows that stability sets normative constraints on justification. A political conception that could not be sufficiently stable is unjustified. As Rawls remarks, "however attractive a conception of justice might be on other grounds, it is seriously *defective* if the principles of moral psychology are such that it fails to engender in human beings the requisite desire to act upon it." (TJ:455/399 rev., my emphasis) Put it another way, it "is imposed as a condition on a reasonable conception of political justice."[3] Furthermore, it is also crucial to political liberalism because the possibility of political liberalism hinges upon a satisfactory answer to the question of "how can the values of the special domain of the political—the values of a subdomain of the realm of all values—normally outweigh whatever values may conflict with them?" (PL:139) We can readily translate this into the problem of moral stability: Given the sense of justice as one motive among other desires, how can it normally be granted priority when they conflict? The concerns of stability turn out to be the same as the concerns of

---

[3] Rawls, *Justice as Fairness: a Restatement*, (Cambridge, Mass: Harvard University Press, 2001), p.89, footnote 10.

political liberalism.

We can now note that stability is essential to justification. It has an important bearing on the desirability of a conception of justice. This claim is however contestable even within Rawls's own theory. In *A Theory of Justice*, Rawls gives readers an impression that stability, though important, is merely concerned with the problem of feasibility after the conception of justice has been worked out in the original position. Similarly, the later Rawls reminds us that his argument for justice as fairness should be viewed as divided into two stages. In the first stage, the parties in the original position select principles without taking the special psychologies and their conceptions of the good into account. The problem of stability arises only at the second stage where its task is to check if justice as fairness is a feasible conception when people have full information about their conceptions of the good.

This division of stages seems to imply that no matter how important stability may be, it has no effect on the justifiability of principles of justice. There is a sharp distinction between desirability and feasibility. The real work of justification takes place in the original position where the principles of justice are chosen by rational parties. What is left for stability is to ensure that the chosen principles are workable and endurable. Freeman, for example, describes the aim of stability as that "assuming we have accounts of the *correct conception of justice*, and of the institutions needed to realize it, how are we to motivate rational persons effectively so that they affirm and support these institutions and the conception of

justice that informs them?"[4] In this case, the arguments for stability add nothing to the correct conception of justice from the point of view of justification. Worse still, given that the right principles have been established in the original position, any change of the first stage argument because of the pressure of stability would be regarded as a trade-off between desirability and feasibility. This explains why Kukathas and Pettit ascribe Rawls's change to political liberalism to his "increasing reliance on the *feasibility* arguments which dominate Part III of his book, and in the corresponding down-playing of considerations of *desirability*."[5]

I believe that this popular reading is wrong. For Rawls, an unstable conception is not only infeasible, but also unjustified. In *Political Liberalism*, for example, he reminds us that the argument in the first stage is only provisionally on hand. "The argument for the principles of justice is not *complete* until the principles selected in the first part are shown in the second part to be sufficiently stable." (PL:141, footnote 7, my emphasis) Moreover, being unstable, "it is not a satisfactory political conception of justice and it must be in some way revised." (PL:141) It is therefore incorrect for critics to say that stability has no bearing on determining a correct conception of justice. It actually plays an important role in defining the reasonableness of principles of justice.

I recognize that Rawls's argument for justice as fairness has in fact experienced a shift from the first stage to the second stage as stability becomes a more salient issue in his search for an overlapping consensus in response to the

---

[4] Freeman, "Congruence and the Good of Justice," p.280, my emphasis.

[5] Kukathas & Pettit, *Rawls: A Theory of Justice and its Critics* (Cambridge: Polity Press, 1990), p.142.

challenge of reasonable pluralism. But if we pay more attention to Rawls's articulation, we will find that the concern of stability has already been central to Rawls's argument in *A Theory of Justice*. In dealing with the problem of envy, for instance, he has said that if a conception of justice derived from the first stage arouses and encourages envy to such an extent that the social system becomes unstable, "*the adoption of the conception of justice must be reconsidered.*" (TJ:531/465 rev., my emphasis) This indicates that although the substantive principles of justice must be worked out independent of people's conceptions of the good, the test of stability has the final appeal to determine which conception of justice is fully justified. If a conception of justice lacks stability, it should either be revised, or given up. As the later Rawls says:

> What if it turns out that the principles of justice cannot gain the support of reasonable doctrines, so that the case for stability fails? Justice as fairness as we have stated it is then in difficulty. We should have to see whether *acceptable changes in the principles of justice would achieve stability.* (PL:65-66, my emphasis)

Why should stability play such a fundamental role in political justification? In particular, why should the overridingness of the sense of justice be viewed as a necessary condition for a reasonable conception of justice? What kind of higher moral end, if any, does stability aim to achieve? These questions must be answered if we want to understand Rawls's theory of justice. What we need is a coherent philosophical account which can reasonably explain the moral significance of stability in Rawls's project. In the rest of this chapter, I will first present a conventional interpretation which offers a standard answer to the above questions, and then argue against its plausibility.

## 2 The Conventional Interpretation

The conventional interpretation holds that the aim of moral stability is to maintain social order. Moral stability matters because it is the most effective means to preserve peaceful and harmonious social cooperation without discontinuity and disruption. As long as a social order lasts over time by whatever means, including persuasion or enforcement, a society is deemed to be stable. Neither the nature of this stabilising force nor where it comes from is the primary concern of social stability. Let us call this *social stability*. The sense of justice is just one of those means. Its importance depends upon its actual contribution to social order. If a conception of justice is proven too demanding to motivate people to comply with its requirements, other means should be deployed. From the point of view of social stability, what makes Rawls's idea of stability so distinctive is not that it serves to realize other higher moral ends, but his belief that there is an intimate causal relationship between moral and social stability. Following this interpretation, Rawls seems to regard social stability as a normative constraint on moral justification. Unless a conception of justice is able to generate sufficient motivational support for an enduring social order, it cannot claim to be justified.

This conventional interpretation has been adopted by almost all Rawls's critics. Barry, for instance, suggests that the later Rawls actually shares the same concern with Hobbes. For Rawls's problem of stability is indeed "the problem of social order."[6] He argues that there is nothing new in Rawls's idea of stability since it has long been a central concern of political philosophy, especially at times

---

[6] Barry, "John Rawls and the Search for Stability," *Ethics* 105 (1995), p.881.

when order is particularly problematic. Rawls shares with Hobbes and all modern natural law theorists a preoccupation with the problem of how people with conflicting religious views and conceptions of the good can live together in peace. They aim to find minimum rules of social interaction that all can reasonably accept. Where Rawls differs from Hobbes is over the means to achieve social order. Barry thus says:

> Formally, Rawls's solution is the same as Hobbes's: that people should retain their differing ends (or conceptions of the good) but reach agreement on certain ideas about what justice requires. Where he departs from Hobbes is in the motivation he seeks for adhering to the dictates of justice.[7]

Kukathas and Pettit also interpret stability as an attempt to find a solution to the problem of how to maintain an enduring social unity in a pluralist liberal society. They hold that the core of Rawls's political liberalism is peace and order. In this respect, "it gives Rawls's politics a decidedly Hobbesian flavour, since he now ties his conception of justice, not to autonomy or individuality, but *order*."[8] In a similar vein, Hampton suggests that the aim of stability is to define a modus vivendi for achieving peaceful cooperation in a community of people with conflicting conceptions of the good. "The public, neutral justification of the project is one that makes it the creation of peace and stability at the lowest political cost, and this is *a Hobbesian justification*."[9] For Hampton, the difference between Rawls and Hobbes is their different approaches to maintain peace. While

---

[7] Barry, "John Rawls and the Search for Stability," p.881.

[8] Kukathas & Pettit, *Rawls: A Theory of Justice and Its Critics*, p.140.

[9] Jean Hampton, "Should Political Liberalism Be Done without Metaphysics?" *Ethics* 99 (July 1989), p.807, my emphasis.

Hobbes appeals to an absolute sovereign, Rawls rests his hope on citizens' sense of justice and an overlapping consensus. An overlapping consensus is more favourable than a modus vivendi simply because the former can achieve peace in a less costly way. Even Habermas, who regards his disagreement with Rawls as one of familial dispute, takes the problem at issue simply as a concern for social stability that "expresses the functional contribution that the theory of justice can make to the peaceful institutionalization of social cooperation."[10] Lastly, Freeman, who shows great sympathy for Rawls's idea of stability, argues that if justice as fairness fails to motivate citizens to act as just institutions demand, "then just social order is unstable and for this reason utopian."[11]

Despite their differences in other respects, the critics seem to share the same view that stability is solely concerned with a practical issue of social order. In this sense, the motivational priority of the sense of justice is merely instrumentally valuable to maintaining social order. Its importance is judged by its degree of contribution to achieving this goal. Hampton thus concludes that "because Rawls's justification of the project of developing an overlapping consensus is instrumental, then no matter what turns out to be required for stability, his project is, and will always be, Hobbesian."[12] It is Hobbesian because the first subject of political philosophy is concerned with social order. Feasibility is prior to desirability.

It is by no means surprising that most critics accept this conventional

---

[10] Habermas, "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," *The Journal of Philosophy* 92 (1995), p.121.

[11] Freeman, "Congruence and the Good of Justice," p.280.

[12] Hampton, "Should Political Liberalism Be Done without Metaphysics?", p.806.

interpretation. For Rawls has seldom made a clear distinction between moral stability and social stability. On a number of occasions he even appears to suggest that the sole concern of stability is about how to maintain an enduring social unity in a pluralist society. Textual evidences can be found throughout Rawls's earlier and later works to support this interpretation. For example, in explaining the structure of justice as fairness, Rawls tells us that the main purpose of the third part of *A Theory of Justice* is to check "if justice as fairness is a feasible conception. This forces us to raise the question of stability and whether the right and the good as defined are congruent." (TJ:580/509 rev.) It means that congruence is required simply because justice as fairness needs to confirm itself to be a workable conception. In *Political Liberalism*, Rawls defines the problem of political liberalism as how it is possible that there may exist over time a *just* and *stable* society of free and equal citizens profoundly divided by reasonable though incompatible religious, philosophical, and moral doctrines. Again, justice and stability are two separate issues. Justice is one thing while stability is another. They do not imply each other.

If so, why should Rawls take social stability seriously? Critics suggest that the answer lies in Rawls's understanding of political philosophy. A general line of reasoning can be sketched as follows. First of all, the main task of political philosophy is to work out a set of basic principles as a common basis to regulate social cooperation and arbitrate conflicting claims. These principles assign rights and duties, and determine citizens' appropriate distribution of benefits and burdens. However, what kind of principles should be adopted is relative to the

historical context of a society. "The aims of political philosophy depend on the society it addresses."[13] In the case of modern liberal democracies the major task of political philosophy is to find a solution to resolve the divisive social and political conflicts arising from the circumstances of justice and the permanent existence of reasonable pluralism. The fact of pluralism has put the problem of social order in the forefront. Without the oppressive use of state power, no comprehensive conception of the good could be accepted as the basis of social unity by free and equal citizens. It naturally follows that finding a neutral political conception of justice that could be the focus of an overlapping consensus is the only way to secure social stability without resorting to coercive force.

Bearing this background in mind, it is not difficult to explain why moral justification should take social stability into account. A conception of justice should not only be desirable from a moral point of view, but also be feasible from a practical perspective. Political philosophy is not intended to be utopian. No matter how perfectly just a principle is, it should be revised or even abandoned if it is incapable of generating sufficient power to guarantee a peaceful and harmonious social order over time. Political justification must be realistic and sensitive to social condition. Therefore, the concern of social order partly determines the acceptability of a conception of justice. Kukathas and Pettit thus suggest that in Rawls's later project, the pursuit of desirability has been subordinated to the consideration of social order. "All consideration of what principles are desirable is framed in the context of the question of what principles

---

[13] Rawls, "The Idea of an Overlapping Consensus," in *Collected Papers*, ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.421.

are most reasonable or feasible for 'us' (i.e., most likely to bring stability)."[14]

If this interpretation is correct, justice as fairness has undergone a fundamental shift from the early Kantian liberalism to the later Hobbesian pragmatism. The worry of social order in the era of pluralism has urged Rawls to re-think the proper role and function of political philosophy. The task of political philosophy is no longer perceived as constructing an ideal conception of justice from an impartial, a-historical and universal perspective, and then applying it to the non-ideal world. Rather, it must confront the challenge of our time. Thus, the nature of Rawls's principles of justice has changed. For instance, freedom and equality are not regarded as valuable in themselves. On the contrary, as Hampton contends, "the ideals of 'fairness' and 'equality' are taken to be instrumentally necessary to the achievement of a stable cooperative society."[15] Each citizen should be treated fairly and accorded the equal protection of basic liberties because this is taken to be necessary for a stable and lasting system of cooperation. The search for social stability overshadows all other liberal values. It becomes the first virtue of political philosophy in the modern democratic circumstance. Its importance arises "from divisive political conflict and the need to settle the problem of order."[16]

## 3 Moral Motivation and Social Order

The conventional interpretation above offers a plausible account of Rawls's

---

[14] Kukathas and Pettit, *Rawls: A Theory of Justice and Its Critics*, p.143.

[15] Hampton, "Should Political Philosophy Be Done without Metaphysics?" p.806.

[16] Rawls, *Justice as Fairness: a Restatement*, p.1.

claim that the problem of stability is fundamental not only to his theory, but also to political philosophy in general. It contains three general claims.

> 1. Social stability matters. Finding ways to maintain a peaceful and enduring social order is the first and foremost task of political philosophy of modern society characterized by reasonable pluralism.
>
> 2. Moral stability is the most effective means to secure social stability.
>
> 3. Social stability is a necessary condition of moral justification.

In this and the following sections, I shall point out that this argument is untenable. This section will focus on (2) and argue that there are no grounds for attributing to Rawls the claim that moral stability arising from the overridingness of the sense of justice and consequent upon an overlapping consensus is the most effective means to attain social stability. In the next section, I will challenge (3) by showing that even if (2) is right, it does not follow that social stability is a necessary condition for political justification. For doing this is not only undesirable, but also conceptually absurd. If my arguments are sound, then no matter how important social stability is, it should have no bearing on the justifiability of justice. The conventional interpretation is flawed and we must look for an alternative argument to make sense of the moral significance of stability.

We now start with (2). Given that social order is what moral stability ultimately aims at, the conventional interpretation can only be sustained on conditions that (a) a majority of citizens will *actually* give regulative priority to

the sense of justice over their conceptions of the good; (b) the fact of overridingness of moral motivation alone will *sufficiently* account for an enduring and peaceful social scheme of cooperation. Applying these conditions to Rawls's political liberalism, it means that social stability depends on moral stability which in turn hinges on whether justice as fairness can be the focus of an overlapping consensus. Only when these conditions are actually met, can the link between the priority of the sense of justice and social order be connected. Social stability is expected to resolve an empirical problem of social order. If the overridingness of the sense of justice is merely a political ideal, the Hobbesians may dismiss Rawls's proposal as useless to confront the challenge of pluralism. Under certain circumstances, an absolute sovereign or an effective system of coercive power may prove to be more appropriate to maintain peace and order. It is not enough for Rawls to say that his proposal is the most effective or most reasonable one in an ideal situation. If the challenge of social stability originates from a practical concern for social order in modern pluralistic society, then moral stability must empirically demonstrate its ability to resolve the problem.

Let us examine condition (a) first. We know that stability depends upon the overridingness of the sense of justice which in turn relies on an overlapping consensus of a political conception of justice. In such a consensus, the reasonable doctrines endorse the political conception, each from its own point of view. "Stability is possible when the doctrines making up the consensus are affirmed by society's politically active citizens and the requirements of justice are not too much in conflict with citizens' essential interests as formed and encouraged by their social arrangements." (PL:134) Rawls's main idea is that since justice as fairness is presented as a freestanding conception without reference to any

comprehensive doctrines, citizens can therefore have sufficient motive to endorse the political conception derived from their own religious, philosophical and metaphysical beliefs. The relation between political values and non-political values "is left to citizens individually—as part of liberty of conscience—to settle how they think the values of the political domain are related to other values in their comprehensive doctrine." (PL:140) Put another way, Rawls expects that Christians, Kantians, utilitarians, Platonists, among others, can converge on his two principles of justice. In case of citizens' non-political values conflict with the requirements of justice as fairness, the holders of these doctrines are still willing to subordinate their interests to the political conception. Moreover, their willingness to do so stems from their sense of justice. This is what a Rawlsian conception of stability requires. For Rawls, then, the success of stability determines the fate of political liberalism because the latter depends on a satisfactory answer to the question of "how can the values of the special domain of the political—the values of a subdomain of the realm of all values—normally outweigh whatever values may conflict with them?" (PL:139) Rawls's challenge is to offer evidence to affirm this claim.

To warrant this kind of consensus, it is not enough that justice as fairness is derived from the public political culture without appealing to, or presupposing any comprehensive doctrines. For this is just one end of political liberalism. Another end is that it must show how each reasonable citizen from their comprehensive perspective *individually* accepts the priority of justice over their conceptions of the good. In such a consensus, the reasonable doctrines endorse the political conception, each from its own point of view. What Rawls should do is to investigate the content of different comprehensive doctrines prevalent in a

democratic society and then to argue from within such doctrines that they have sufficient grounds to support the requirements of justice as fairness. Such a consensus cannot be determined by Rawls's own Kantian theory, which is only one among many comprehensive views. Nor can Rawls claim that as a philosopher, he can reach a conclusion from an impartial point of view by taking all comprehensive perspectives into account. He must leave the decision to each citizen. Surprisingly, Rawls has never made such an attempt. He does not conduct any empirical survey to show that justice as fairness is the most reasonable candidate for an overlapping consensus. He even admits that his project is just a kind of uncertain speculation. As he puts it, "whether justice as fairness (or some similar view) can gain the support of an overlapping consensus so defined is a speculative question." (PL:15)

Rawls does not explain why he sets this empirical approach aside. One reason may be that this approach is too difficult to realize. It is a daunting, if not impossible, task for a political philosopher to examine the moral and philosophical premises of every comprehensive doctrine and then determine whether they can converge on a political conception of justice. Besides, another factor adds difficulty to this approach. In principle, citizens are identified with regard to the comprehensive doctrines that they hold. But in fact, most people are not philosophers. Their systems of belief are always inarticulate, mixed, inconsistent and unstable. Consequently, their motivation to act can be intermingled with different reasons. In this case, as Hill worries, "even winning the allegiance of the major religions and philosophical theories (for justice as fairness) would still not ensure stability; the more or less doctrineless folk need to be convinced as well, and they are already averse to philosophical systems of

ideas."[17]

If my analysis is plausible, the first condition of social stability is in doubt because no empirical evidence is available to confirm the success of an overlapping consensus. Someone may argue that regardless of the lack of empirical support, it is however certain that justice as fairness as a political conception is more stable than any alternative comprehensive doctrines. This defence is insufficient at least for two reasons. First, if the idea of the political conception gives justice as fairness an advantage over other competing theories of justice in this regard, other theories can follow Rawls's lead and present themselves as a political conception as well. For instance, a libertarian political conception is conceivable. Rawls actually acknowledges this possibility. Therefore, the philosophical debate among different conceptions of justice cannot be resolved simply by appealing to the idea of a freestanding political conception. Second, for the sake of argument, even if we agree that justice as fairness is more stable than most comprehensive political doctrines, this does not mean that it is sufficiently stable in the Rawlsian sense which requires that a conception of justice can generate the sense of justice from each citizen's subjective motivational set to grant priority to the political values. It may still be a modus vivendi.

In order to tackle this difficulty, Rawls attempts to use several model cases to demonstrate that these comprehensive doctrines would endorse his political conception as a balance of reasons as seen within each citizen's comprehensive

---

[17] Thomas Hill, "The Stability Problem in *Political Liberalism*," *Pacific Philosophical Quarterly* 75 (1994), p.342.

doctrine rather than a pragmatic compromise compelled by circumstances. These cases include the religious doctrines that accept the principle of toleration, the liberalisms of Kant and Mill, a mixed partially comprehensive doctrine and utilitarianism. (PL:145, 170). The use of model cases does not aim to offer an empirical proof. But Rawls does hope to demonstrate how an overlapping consensus may proceed. [18] Let us take utilitarianism as an example and see if this demonstration is successful. Rawls's argument is as follows:

> This utilitarianism supports the political conception for such reasons as our limited knowledge of social institutions generally and on our knowledge about ongoing circumstances ... These and other reasons may lead the utilitarian to think a political conception of justice liberal in content a satisfactory, perhaps even the best, workable approximation to what the principle of utility, all things tallied up, would require. (PL:170)

We can note that the main reason for a utilitarian to accept political liberalism is that the liberal arrangement of the basic structure is approximately the best means to maximize utility. He still regards the principle of utility as the highest moral principle. As Rawls acknowledges in *A Theory of Justice*, from a utilitarian point of view, their acceptance of the priority of the right is based on a contingent fact that "under the conditions of civilized society there is great social utility in following them for the most part and in permitting violations only under exceptional circumstances." (TJ:28/25 rev.) But when circumstances are different, utilitarianism may favour another political arrangement. Justice as fairness is only instrumentally valuable to the maximization of total utility. Thus, "while the contract doctrine accepts our convictions about the priority of justice as on the

---

[18] I will give a more detailed examination of this argument in Chapter 5.

whole sound, utilitarianism seeks to account for them as a socially useful illusion."(TJ:28/25 rev.) Suppose Rawls does not change his attitude toward utilitarianism in this regard, it is hard for him to consistently hold that in an overlapping consensus, "no one accepts the political conception driven by political compromise." (PL:171) Utilitarianism is an example to show exactly that a reasonable doctrine may not have the right reason to accept the overridingness of the political values. It merely views Rawls's principles as a modus vivendi. Utilitarianism is however not an exception. If Rawls's idea of an overlapping consensus succeeds, it seems that most fundamental and controversial disputes in political philosophy will be resolved or set aside to the non-political sphere because the major competing theories of justice are all presumed to be able to endorse political liberalism as a higher-order principle. Nevertheless, if it is a permanent fact that none of the reasonable comprehensive doctrines can be affirmed by all citizens in a modern democratic society, this situation should apply to the sphere of the right as well. History tells us that citizens disagree about what is right as bitterly as about what is good.

Furthermore, even if condition (a) is met, moral stability is still not enough to ensure a lasting and peaceful social scheme of cooperation. This is because condition (b), which holds that the fact of the motivational priority of the sense of justice alone will sufficiently account for an enduring and peaceful social scheme of cooperation, is untenable if stability is interpreted as the Hobbesian concern for social order.

Recall that Rawls's argument for stability depends upon the overridingness of the sense of justice. The idea of an overlapping consensus is designed to

achieve this goal through gaining the reasoned support of citizens who affirm reasonable, though conflicting, comprehensive doctrines. Once a consensus is realized, social stability will necessarily follow. Rawls seems to believe that citizens' effective desire to act justly is indispensable to an enduring social order. But the connection between moral and social stability is not necessary. Undoubtedly, a wide-spread allegiance to the liberal principles of justice can positively contribute to a stable social cooperation. However, moral motivation is just one among many factors to affect social stability. The sustainability of a political regime is also subject to other contingent factors, such as the degree of economic development, religious and political culture, division of social class, and racial and ethnic relations. It is a commonplace in political sociology that maintaining a peaceful and harmonious social order requires a state to take particular historical conditions into consideration. Different societies face different problems, and require different institutions and policies. Reasoned and philosophical arguments alone are insufficient to secure a stable order. As Hill rightly questions, "the factors which stabilize various societies may in fact have relatively little to do with the systems of ideas that they espouse, and more to do with habit, reinforcement, and blind emotional attachment."[19] Klosko makes a similar remark:

> [Rawls's] entire treatment of political stability is hampered by his failure to examine different factors that contribute to, or weaken, this. His emphasis on moral stability above all other factors would strike most political sociologists as, at best, an unusual claim, lacking either empirical or philosophical

---

[19] Hill, "The Stability Problem in Political Liberalism," p.342.

support."[20]

Rawls probably believes that in a well-ordered society the only factor that would affect social stability is citizens' moral motivation. This assumption is, however, ungrounded. To conclude, even if social stability is the prime concern of political philosophy, Rawls's political liberalism fails to show that the majority of citizens would actually give priority to the sense of justice over their conceptions of the good; it also fails to establish the necessary link between moral stability and social stability. Furthermore, for the sake of argument, even if (2) is valid, I believe that the third claim, which holds that social stability is constitutive of justification, is wrong. This is the main concern of the next section.

## 4 Social Order and Justifiability

Critics have complained that the later Rawls has changed from a Kantian to a Hobbesian by accepting the primacy of social stability in justification. They rarely disagree with Rawls over the source of stabilising force. What they object to is the view that a conception of justice can claim to be justified if and only if it establishes a long-lasting and peaceful scheme of cooperation. For this amounts to saying that political liberalism is justified mainly because it is the most effective means to attain social stability. The primacy of social order defines the character of political liberalism. The dispute at issue is whether Rawls truly adopts a Hobbesian model of political justification, and if he does, whether it is desirable for him to make such a shift. To answer these questions, a brief account of

---

[20] George Klosko, "Rawls's Argument from Political Stability," *Columbia Law Review* 94, (1994), pp.1891-92.

Hobbes's view is warranted before we probe into Rawls's own argument.

Hobbes's contract theory is a model of justice as mutual advantage. Hobbes portrays human beings as profoundly self-interested individuals.[21] People's sole motivation to act is their interest in pursuing and protecting their self interest. Furthermore, he takes it for granted that all rational beings have a common dominant end to preserve life and avoid violent death. However, in a lawless state of nature conflicts of interest and the struggle for power define the human condition. It is characterized by the war of every one against every one. Without a state to enforce laws and constrain individuals' behaviour, individuals enjoy "natural rights" to use all means to protect their lives and to do whatever they wish to further their interest. In the state of nature, people are living in a constant struggle for survival. From this perspective, peace and social order are paramount. They recognize that it is irrational to stay in the state of nature. For the unconstrained pursuit of their own interest is bound to lead to conflict. The key problem, in Hobbes's view, is: under what conditions can rational self-interested individuals come to respect and trust one another, honour and comply with a set of principles so that their long-term interest in security and social order can be upheld and sustained? [22] This set of principles defines what justice is. As Freeman aptly puts it, "a just society for Hobbes is nearly identifiable with a stable social order. He conceives of justice as people's mutual compliance with

---

[21] Hobbes does actually acknowledge that there are some noble characters who give priority to justice even over their lives. But they are such a small minority that they do not count politically. I am indebted to John Charvet for this point.

[22] Hobbes's view can be seen in *Leviathan* ed. Richard Tuck (Cambridge: Cambridge University Press, 1991).

the norms and institutions needed to establish peaceful social cooperation."[23] The concern of social stability constrains what kind of substantive political principles would be agreed in the first place.

The next question is how this conception of justice can be worked out and complied with by self-interested individuals. The general Hobbesian idea is essentially that individuals should willingly surrender their rights enjoyed in the state of nature and come to agree on a set of rules of cooperation that reflects the balance of bargaining power of different parties. It is rational for them to do so because complying with the rules can bring more advantages to everyone. This accounts for the term "justice as mutual advantage."[24] Moreover, the rules are the result of the consent of individuals from their personal point of view. For Hobbes, people are not required to give up their existing self-interest for the sake of justice derived from some impersonal grounds. For in essence, self-interest and justice are different sides of the same coin. What is just is derived from what is good judging from their rational point of view. Strictly speaking, there is no conflict between justice and self-interest. They refrain from breaking the rules because doing so is more conducive to advancing their long-term interest. As Barry succinctly puts it, for justice as mutual advantage, "a set of rules is just if general compliance with the rules would be more advantageous to everybody (in terms of each person's conception of the good) than the alternative of a 'state of nature' in which everybody pursued their conception of the good without any constraints."[25]

---

[23] Freeman, "Congruence and the Good of Justice," p.278.

[24] It should be noted that justice as mutual advantage refers to a general conception akin to Hobbes's thought. I am not saying that Hobbes exactly holds this view. The most important work of this approach is David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986).

[25] Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), p.46.

In other words, justice is a result of a practical compromise, or a modus vivendi among rational self-interested parties. As a result, social order is contingent on circumstances favourable to convergence of interests. Once the distribution of political power shifts, the existing conception of justice is likely to be upset and re-negotiations are thereby required so as to reach new terms of cooperation. In order to resolve this assurance problem permanently, Hobbes holds that an unconstrained sovereign empowered to use the monopoly of coercive political force is needed to guarantee a peaceful social order.

Given these defining features of justice as mutual advantage, Rawls's conception of stability can correspondingly be interpreted as Hobbesian in two different ways. It is Hobbesian either because Rawls defines the primary goal of justice as equivalent to maintaining peace and social order, or because justice as fairness, as a matter of fact, turns out to be a modus vivendi. These two questions are conceptually distinct from each other. We can accept the primacy of social order while believing that a non-modus vivendi social cooperation is achievable, and *vice versa*. In the following I will refute both claims ascribed to Rawls.

I will first examine the problem of modus vivendi. If we look at Rawls's characterisation of stability carefully, it is clear that he has no intention to present justice as fairness as a contingent balance of power among essentially conflicting interests. Recall that the necessary and sufficient condition of stability rests on the overridingness of the sense of justice. The reason for people to adhere to political principles stems from their effective moral motivation rather than rationally calculated self-interest. "Stability is secured by sufficient motivation of the appropriate kind acquired under just institutions." (PL:142-43) A fundamental

difference between Hobbes and Rawls is their understanding of human nature. Hobbes believes that human beings are essentially egoistic. Their adherence to the rules is regarded as rational only to the extent that it can be anticipated to advance their self-interest. Rawls, on the contrary, believes that agents have a capacity and a desire to act on principles that can be reasonably accepted by free and equal citizens in a fair and equal-footing condition. Justice as fairness is a moral conception affirmed on moral grounds. Its justification is based on moral reasons rather than on the fortunate convergence of conflicting interests. As a result, "those who affirm the various views supporting the political conception will not withdraw their support of it should the relative strength of their view in society increase and eventually become dominant." (PL:148)

Critics may continue to argue that regardless of Rawls's effort to distance his project from a modus vivendi, he fails to deliver his promise, to wit, a stable well-ordered society secured by the argument of congruence between the right and the good presented in *A Theory of Justice,* or by the argument of overlapping consensus presented in *Political Liberalism.* Or they may argue that Rawls's understanding of moral psychology is simply unrealistic. Human nature is too crooked to do the right thing. These are vital challenges that I will take up in the coming chapters. For the moment it suffices that from Rawls's own point of view, a modus vivendi is neither the aim of his project nor a necessary result of his philosophical construction.

We now turn to the second charge, namely whether the concern for stability has made Rawls a Hobbesian by accepting the primacy of social order. Before answering this question, I will first offer an account of why Rawls's project of

political liberalism has been widely perceived in this way as the conventional interpretation does, and then point out why this interpretation is unacceptable.

According to Rawls, political liberalism aims to deal with two fundamental questions about political justice in a pluralistic society. The first concerns how to find the most appropriate conception of justice to regulate social cooperation between free and equal citizens regarded as fully cooperating members of society. The second concerns how this conception can gain the allegiance of citizens and therefore ensure the priority of political values over non-political ones. Combining the two questions together, the central question of political liberalism is: "how is it possible for there to exist over time a *just and stable* society of free and equal citizens, who remain profoundly divided by reasonable religious, philosophical, and moral doctrines?" (PL:4, my emphasis)

This indicates that Rawls purports to achieve two goals simultaneously in his project, to wit, a conception of justice must be right and feasible. The desirability of a conception of justice is distinct from, and independent of, its feasibility. The question of desirability is concerned with the moral grounds of a conception of justice. Rawls's answer is that a reasonable conception of justice must be based on an idea of society as a fair system of cooperation, together with a conception of moral persons as free and equal with two higher-order interests in developing their moral capacities for a sense of justice and a conception of the good. In addition, the principles must match our considered judgments in reflective equilibrium. By contrast, the question of feasibility is concerned with how a conception of justice can effectively and smoothly apply to the basic structure of society and generate its own force to last over time. It is a practical matter irrelevant to justifiability of

a conception of justice. Thus whether a political principle can be morally justified is *conceptually* different from whether it can be socially stable. It is entirely possible that a principle can be regarded as just from a moral point of view while failing to be endurable owing to empirical constraints of social and political conditions. They do not entail each other.

In order to realize these two goals, Rawls attempts to present justice as fairness in two stages. At the first stage, two principles of justice are derived from the original position behind the veil of ignorance. The design of the original position expresses Rawls's ideal of society as a fair system of cooperation between free and equal persons. It represents a neutral and freestanding point of view that does not presuppose or appeal to any conception of the good. The problem of stability only comes about at the second stage where the veil is lifted and the principles of justice have already been on hand. Rawls says that this division of labour ensures that the content of justice will not be affected by any particular comprehensive doctrines. It will not be an outcome of political compromise or bargaining. The question of overlapping consensus arises in the second stage because it is designed to resolve a practical issue of social stability.

Following this account, whether a conception of justice can be the focus of an overlapping consensus has no impact on its justifiability. For its function is to deal with feasibility rather than desirability which has already been settled in the first stage. No matter how important social stability is from the practical point of view, it is irrelevant to the justifiability of principles of justice. In this sense, the first stage is normatively prior to the second stage. Habermas has made a fairly incisive comment on this point, which is worth quoting in length. He says:

> Because Rawls situates the "question of stability" in the foreground, the overlapping consensus merely expresses the *functional* contribution that the theory of justice can make to the peaceful institutionalization of social cooperation; but in this the intrinsic value of a *justified* theory must already be presupposed...The overlapping consensus would then be merely an index of the utility, and no longer a confirmation of the correctness of the theory; *it would no longer be of interest from the point of view of acceptability, and hence validity, but only from that of acceptance, that is, of securing social stability.*[26]

If Habermas's observation is correct, the overlapping consensus and social stability have no relevance to the justifiability of justice as fairness. However, this position is not what Rawls actually holds. In reply to Habermas's doubt about whether the overlapping consensus can add anything to the acceptability of a political conception, Rawls states:

> Only when there is a reasonable overlapping consensus can political society's political conception of justice be publicly—though never finally—justified ... There is, then, no public justification for political society without a reasonable overlapping consensus, and such a justification also connects with the ideas of stability for the right reasons as well as of legitimacy. (PL:388-89)

This reply has squarely refuted Habermas's interpretation. What Rawls actually holds is that an overlapping consensus is a prerequisite for public justification. If a conception of justice is shown to be unstable, it is not only infeasible, but also unjustifiable. It clearly indicates that the supposed distinction

---

[26] Habermas, "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," pp.121-22, my emphasis.

between desirability and feasibility is not held by Rawls. Unless a political conception is sufficiently stable, the argument is incomplete and "it must be in some way revised." (PL:141)

If we follow this account, the concern of stability has fundamentally changed the justificatory structure of justice as fairness. The arguments presented in the original position are no longer self-sufficient to provide full justification for his two principles of justice. It must wait for confirmation of the overlapping consensus at the second stage. The focus has shifted from the first stage to the second stage. Why does Rawls make such a big move? The conventional interpretation immediately comes on the scene and offers a ready answer: Rawls has changed from a Kantian to a Hobbesian. This is in a sense understandable. It is well known that the original position expresses a Kantian account of human beings as free and equal moral agents with a higher order interest in realising their true nature through fair social cooperation. Now if an overlapping consensus sets limits on the principles of justice derived from the original position, the only explanation would be that Rawls has taken the issue of social order as the primary concern of social justice. Desirability succumbs to feasibility.

If this conventional interpretation is correct, Rawls needs to pay a heavy moral cost for his Hobbesian turn. The reason is this: in case there is an irreconcilable conflict between desirability and feasibility, the former cannot but make a concession. For instance, imagine that the difference principle is shown to be unable to gain an overlapping consensus and is therefore unstable in a society. What we should do, Rawls would suggest, is not to insist on the correctness of the principle, but rather to "see whether acceptable changes in the principles of justice

would achieve stability. (PL:66) This is a concession because justifiability of the difference principle has to be adjusted or even sacrificed for the sake of a practical concern for social order. Feasibility sets a constraint on what justice should be like. But a socially stable conception of justice may be morally unjustified. In an individualist capitalist society, for example, the difference principle could be less likely to be the focus of a consensus if Nozick's libertarianism is more prevalent among reasonable comprehensive doctrines. In this case, libertarianism can claim to be more justifiable than justice as fairness because it has a better chance to achieve social stability. Rawls would therefore not be able to reject Nozick's entitlement theory by arguing that his conception of moral personality and of social cooperation is morally unacceptable.

This concession is unacceptable because the quest for social stability has overshadowed other moral considerations. We can ask whether it is worthwhile to compromise the integrity of moral principle for a practical constraint, and whether it is reasonable to lower the moral standard to cater for social order. If the primary task of political philosophy is to justify a morally defensible and attractive conception of justice, bringing social stability into justification will hamper rather than strengthen the normative character of political philosophy. As Arneson remarks, "we cannot decide on appropriate proxy measures for the in practice unmeasurable qualities we really care about until we decide what we really care about. At this stage in our inquiry the appeal to the constraints of feasibility is premature. "[27] Furthermore, even if a political theory can be designed to ensure a

---

[27] Richard Arneson, "Responsibility, Neutrality, and Political Liberalism," (1996), typescript. Similar criticisms can also be found in David Copp, "Pluralism and Stability in Liberal Theory," *The Journal of Political Philosophy*, 4 (1993): 204-05. Chandran Kukathas and Philip Pettit, *Rawls:*

lasting social order, it would still be undesirable if its content is unjust in the first place. As Freeman aptly observes, "by itself a stable social order, however rational it may be, can be of little moral consequence if it does not rectify but only perpetuates gross injustice."[28] To conclude, if the conventional interpretation of Rawls's conception of stability is correct, then political liberalism may exactly face the criticism Freeman raises. Taking stability as a constraint on justification is a setback rather than an improvement in Rawls's theory.

Someone may try to rescue Rawls from this predicament by saying that the discrepancy between the first and the second stage would never happen. A political conception of justice derived from the original position would necessarily lead to an overlapping consensus after the veil is lifted. In other words, the first stage entails the second stage. But if so, there is no point in Rawls including the second stage in the process of justification. The issue of an overlapping consensus becomes redundant from the point of view of justification. Moreover, as shown above, the reasons to support the principles of justice at the first stage is qualitatively different from the reasons to assure social stability at the second stage. Differing from the original position, the argument that takes place in the second stage, as Habermas points out, "refers not to the fictional citizens of a justice about whom statements are made within the theory but to real citizens of flesh and blood. The theory, therefore, must leave the outcome of such a test of acceptability *undetermined*."[29] It is undetermined because the outcome depends

---

*A Theory of Justice and its Critics*, pp.142-50.

[28] Freeman, "Congruence and the Good of Justice," pp.278-79.

[29] Habermas, "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," p.121, my emphasis.

upon the empirical conditions which are out of Rawls's control. There is no ground for Rawls to guarantee an antecedent linkage between the two stages.

Furthermore, a more serious critique, which has already been implicated in the above comment, is that the conventional interpretation will make Rawls's account of the relation between stability and justification conceptually absurd. The reason is this. Under the conventional interpretation, Rawls would hold that there is no justification of a conception of justice without social stability. Having a propensity to endure is a constitutive feature of justice. A justifiable conception of justice is by nature stable. However, this connection is not only morally undesirable as argued above, but also conceptually absurd. For justice and social stability are conceptually distinct from each other. They belong to two different categories. If we treat stability as a condition of justice, according to Cohen:

> It would mean that one could not say such entirely intelligible things as "This society is at the moment just, but it is likely to lose that feature very soon: justice is such a fragile achievement"; or "We don't want our society to be just only for the time being: we want its justice to last". It would mean that Plato was conceptually confused when he argued, on empirical grounds, in Book VIII of *The Republic*, that a just society was bound to lose its justice over time.[30]

Cohen's argument looks simple, but fatal to the conventional interpretation of stability. Cohen's point is that treating social stability as a requirement of justice commits a category mistake. In considering fundamental principles of justice,

---

[30] G. A. Cohen, "Rescuing Justice from Constructivism," typescript, unpublished. The prelude of this long article has been independently published entitled "Facts and Principles," *Philosophy and Public Affairs* 31, no.3, (2003), pp.211-45.

social stability should not be taken into account. What justice is should only be argued on moral grounds, but not on empirical pragmatic considerations. This does not mean that the problem of stability is not an important issue in political philosophy. A stable and lasting social order regulated by a conception of justice is undoubtedly essential to an effective social cooperation. But stability matters only at the application level where the fundamental principles have already been justified. Whether a political principle is just in theory should be distinguished from whether it is stable in practice. It is simply wrong to expect that a just principle is necessarily stable. For they are judged by different criteria. This points to a conclusion that the concern for social order should have no impact on the justifiability of justice at all. Only after drawing this distinction can it be sensible for us to say such things as "a political principle is perfectly just though it is fragile and unstable because its requirement may be too demanding or uncongenial to the actual social environment." If the conventional interpretation is valid, Rawls would have made a serious conceptual mistake in the first place, and the third claim of the conventional interpretation mentioned in the previous section will be unsound.

Cohen forces Rawls to face a dilemma here. On the one hand, Rawls claims that social stability is a precondition of the justifiability of a conception of justice; on the other hand, he must acknowledge that under certain circumstances even a just society may be less stable than a less just one. But these two claims cannot be both valid. If we hold the former, then it is by definition true that a just society is stable. This, as Cohen shows, does not make any sense though. Even Rawls on occasion concedes that "a just scheme of cooperation may not be in equilibrium, much less stable." (TJ:496/434 rev.) If Rawls holds the latter, he must then forsake

his claim that stability is a necessary condition of justification. Then his project of political liberalism would collapse. The conventional interpretation does not show promise in resolving this dilemma.

The only way to resolve this dilemma, I believe, is to disassociate moral stability from social stability. We should neither regard moral and social stability as expressing the same concern for a conception of justice, nor view moral stability as a necessary condition of social stability. Rather, they should be regarded as being concerned with different issues. If it can be established that moral stability has its independent agenda that is essential to political justification, it can then allow Rawls to say that political justification must take moral stability into consideration while being consistent in holding that a morally stable conception of justice may sometimes be socially unstable in practice.

Once this distinction is drawn, the conventional interpretation should be rejected and the above criticism against Rawls can be settled. For instance, Rawls's change to political liberalism need not be interpreted as a move from a Kantian conception to a Hobbesian one, and his argument for stability need not be viewed as a compromise between feasibility and desirability. The concern of stability itself expresses a moral standpoint. In short, justice is not sacrificed for the sake of social order. Finally, giving moral stability an independent status will provide us with a new perspective to understand Rawls's philosophical enterprise.

## 5 Conclusion

In this chapter I have presented Rawls's conception of moral stability and sketched out its role in his theory. I have also drawn a distinction between moral

stability and social stability and shown why it is necessary to do so. There is no doubt that for Rawls moral stability is essential to the justification of justice as fairness. The difficulty is to explain why this is so. The most puzzling question is why Rawls sometimes presents the problem as simply a concern for feasibility which is independent of desirability while stressing that it should be imposed as a constraint on justification. It is difficult to reconcile these two conflicting accounts.

To resolve this difficulty, I take pains to articulate and then refute a conventional interpretation adopted by most critics who ascribe the meaning of stability to a Hobbesian concern for social order. I have argued that viewing social stability as a constitutive feature of a conception of justice is not only undesirable, but also conceptually absurd. Facing this consequence, we have two options. The first is to continue to uphold the conventional interpretation while finding other ways to respond to the criticism. I am pessimistic about this approach. The alternative is to search for a new interpretation which can avoid the loopholes of the conventional one and in the meantime offer a more consistent and appealing picture to make sense of the significance of stability. This is the approach that I am going to argue for in the next chapter.

## CHAPTER 2

# THE IMPORTANCE OF MORAL STABILITY

In the last chapter I suggested that moral stability should have its own agenda which is distinct from social stability. This chapter aims to vindicate this claim and argue for the importance of moral stability in political justification. It attempts to explain what this agenda is and why it matters so much in Rawls's theory. In other words, I purport to affirm his claim that stability is fundamental to political philosophy.

The clue to answer this question hinges upon Rawls's account of the motivational priority of the sense of justice. I have previously shown that the priority of moral motivation is the necessary and sufficient condition of stability. A stable conception of justice, by stipulation, should be able to "generate in its members a sufficiently strong sense of justice to counteract tendencies to injustice." (PL:140-41) Furthermore, it depends upon a balance of motives, that is, "the sense of justice that it cultivates and the aims that it encourages must normally win out against propensities toward injustice." (TJ: 454/398 rev.) In other words, the agenda of stability is about how a theory of justice can provide sufficient moral motives for each rational agent to act in accordance with the command of justice. To understand the significance of moral stability, we should explore the nature of this priority claim and uncover its relation with moral justification.

The structure of this chapter is arranged as follows. Section 1 will define the notion of "motivational priority of justice". Section 2 will discuss the idea of a

rational plan of life which defines a person's good and provides a teleological framework for practical reasoning. After that, in Section 3, I argue that Rawls has adopted a reason/motive internalism which, when combined with prudential rationality, will explain why the motivational priority of justice can only be accounted for by showing its regulative status in an agent's rational plan of life. Once this is established, I go one further step in Section 4 to justify the importance of motivational priority by appealing to Korsgaard's idea of the normative question. Finally, in Section 5 and 6, I respond to two criticisms raised by Barry and Sandel respectively, and further demonstrate the distinctive feature of Rawls's account of stability.

## 1 The Motivational Priority of Justice

To begin with, we need to know what the "motivational priority of justice", or "MPJ" for short, exactly means. MPJ is a claim that doing what justice forbids can never be what one has most reason to do.[1] Put it another way, it requires that it is always rational for an agent to have compelling reasons to do what justice commands. Two important points about this claim should first be noted.

First, MPJ is a normative claim about the motivational efficacy of a conception of justice. To support MPJ, a theory of justice needs to provide a general account of human motivational structure, and explain how the moral point of view specified by that theory can have such normative force as to outweigh other desires and command people's compliance. This pertains to practical reason

[1] Here I follow Scheffler's definition of the claim of overridingness. But while Scheffler is concerned with overridingness of morality in general, my concern is of a particular conception of justice. Samuel Scheffler, *Human Morality*, (New York: Oxford University Press, 1992), pp.52-53.

for action. It is a goal that a moral and political theory should aim to achieve. But its possibility largely depends on the attractiveness of the conception of justice in question. Substantive arguments are required if we want to compare two theories in terms of their motivational efficacy.

We should therefore not be misled into thinking that moral stability is determined by how many people *actually* act in accordance with principles of justice in a society. If interpreted in this way, Rawls's whole project, as argued in the last chapter, can hardly be defended. No one can deny that other things being equal, a morally stable conception of justice will make a substantial contribution to an enduring social order. When the principles of justice are widely respected by a majority of citizens, it is indeed a good sign that social cooperation under the existing basic structure may last. But this does not mean that a socially unstable society necessarily reflects a failure of moral stability. For there may be other factors at work. For example, some people are too egoistic or irrational to follow the political principles. If they were rational and reasonable enough, they should have granted the priority of justice over other desires.

My point is that while there is a positive correlation between MPJ and social stability, they should not be regarded as expressing the same concern. Social stability is concerned with a practical issue of the feasibility of a conception of justice, while moral stability is concerned with the normative priority of moral consideration in practical reasoning. They are conceptually different from each other. Rawls, however, seems to disagree with me. For example, he says,

> Since a well-ordered society endures over time, its conception of justice is presumably stable: that is, when institutions are just

> (as defined by this conception), those taking part in these arrangements acquire the corresponding sense of justice and desire to do their part in maintaining them. (TJ:454/398 rev.)

This suggests that moral stability results in a socially stable society. But this only occurs in an ideal well-ordered society in which the sense of justice is deemed the only factor affecting social order. This is only one of many scenarios. It is possible that a morally stable conception of justice may not lead to a long-lasting social order for non-moral reasons. Moreover, people do not always act in accordance with reason. As Rawls says, "to justify a conception of justice we do not have to contend that everyone, whatever his capacities and desires, has a sufficient reason (as defined by the thin theory) to preserve his sense of justice." (TJ:576/504-505 rev.) It indicates that even if a conception of justice is justified and therefore *rational* for people to respect the claim of MPJ, some people could still act otherwise because of their self-regarding interests. Acceptability is thus distinct from acceptance. No matter how important social stability is, it should not be imposed as a constraint on the desirability of a conception of justice. Doing this will commit a category mistake. I believe that the later Rawls is well aware of the danger of this confusion. That is why he particularly emphasizes that "stability means stability for the right reasons. This implies that the reasons from which citizens act include those given by the account of justice they affirm...which characterizes their effective sense of justice." (PL:xlii) This clearly shows that what Rawls cares about is not social stability *per se*. He wants a particular kind of stability whose stablizing force is solely derived from moral motivation. Put another way, if a socially stable society did not come about from the *right* reasons, Rawls would not deem it as morally stable. So, his conception of stability is closely tied to the justifiability of a conception of justice. It is a property of a

conception of justice.[2] Once the distinction is drawn, many charges against Rawls's conception of stability raised in the previous chapter can now be left aside, and we can focus on the claim of MPJ.

The second point to be made about the formulation of MPJ is the definition of rationality. MPJ stipulates that it is always rational for an agent to do what is just. The criterion of rationality is, however, full of controversy. There are many competing interpretations of rationality. What is certain is that a rational action should not be vaguely described as motivated by some reasons or other. For all intentional actions have reasons to support them in this sense. The definition must be specific and substantive so that it can provide a clear guideline to judge when a claim of MPJ is satisfied. Fortunately, Rawls has offered such an interpretation, namely that an act is rational when it can best promote a person's good in accordance with his rational plan of life which is in turn defined by principles of rational choice and full deliberative rationality. Let us call this *prudential rationality*. As this conception of rationality is essential to Rawls's argument for stability, I will first explain the main ideas of this conception, and then demonstrate its linkage with the claim of MPJ in the following sections.

## 2 The Idea of a Rational Plan of Life

First of all, Rawls assumes that the very nature of human action is intentional and teleological. By this he means that people are normally motivated to act by their goals, desires and plans. They seldom act arbitrarily without any purposes.

---

[2] Rawls, *Justice as Fairness: A Restatement* ed. Erin Kelly (Cambridge, Mass: Harvard University Press, 2001), p.181.

Everyone has a life to lead, and desires to live well. People are fundamentally motivated to act by their conceptions of the good.[3] I take this as the starting point for Rawls's theory. It is related to Rawls's conception of society as a cooperative venture for mutual advantage marked by a conflict as well as by an identity of interests:

> There is an identity of interests since social cooperation makes possible a better life for all than any would have if each were to live solely by his own efforts. There is a conflict of interests since persons are not indifferent as to how the greater benefits produced by their collaboration are distributed, for in order to pursue their ends they each prefer a larger share to a lesser share. (TJ:4/4 rev.)

The major reason for people to participate in social cooperation is that this can better promote their interests and ends. Before they start to negotiate the terms of cooperation in the original position, participants are supposed to have their own conceptions of the good. The primary motive for cooperation is to acquire more primary goods to advance their rational plans of life. They are not supposed to be moved by benevolence or other moral sentiments. If living alone could give them a better chance to have a better life, they would have no reason to enter into social cooperation and abide by rules of justice. In this sense, the good is prior to the right. For without a sufficient incentive to advance their good, the process of searching for a conception of justice will not even start off. This explains the idea of justice as mutual advantage.[4]

---

[3] Of course I am not saying that every agent acts this way in every single action. I think it suffices if this claim can explain most of our actions in normal circumstances.

[4] The later Rawls explains that his theory should have been called "justice as reciprocity" while 'justice as mutual advantage' is left for the Hobbesian model of social justice. PL:16-17.

This model of cooperation, however, immediately poses a threat to Rawls's theory: if participants are primarily motivated by self interest, how can the claim of MPJ be possible? Why should people be moral if their conceptions of the good are in conflict with the demands of justice? This is a big problem. The possible tension between justice and self-interest and the demand of priority for the former sets the background for the discussion of moral stability. Rawls's immediate answer is that "if men's inclination to self-interest makes their vigilance against one another necessary, their public sense of justice makes their secure association together possible...the general desire for justice limits the pursuit of other ends." (TJ:5/4-5 rev.) We can notice that Rawls acknowledges the conflict between self-interest and justice. He has no intention to claim that every person is or should be a moral saint. It is legitimate for people to pursue their interest. What he expects is that the latter can be justified to have priority over the former when the conflict arises. To what extent this claim is sound will be discussed later in my thesis. My point here is that even if the sense of justice is a fundamental desire in one's "subjective motivational set",[5] Rawls does not deny that the desire for one's own good also plays an important role in practical reasoning.

Rawls's challenge is therefore to argue why it is rational for an agent to give priority to justice over his pursuit of the good. This is the central issue of moral stability. To resolve the tension between justice and self-interest, Rawls argues that the sense of justice must be "desirable from the standpoints of rational persons who have them when they assess their situation independently from the

---

[5] This term is drawn from Bernard Williams, "Internal and External Reason," *Moral Luck*, (Cambridge: Cambridge University Press, 1981), pp.101-113.

constraints of justice." (TJ:399/350 rev.) The moral motive is not by default overriding. Its priority has to be justified by showing that acting justly is an important good for rational agents judging from their first-person perspective. Acting justly is not something externally imposed on an agent. Nor does it necessarily contradict our good. Rather, a reasonable conception of justice should be able to show that justice specified by the political principles can be presented as constitutive of our well-being and play a regulative role in our life. Rawls calls this the *congruence* argument. On the one hand, Rawls recognizes the tension between justice and self-interest, and on the other hand strives to vindicate that this tension is not necessarily unsolvable. With a reasonable political morality and a proper account of reasons for rational actions, the tension can be relieved, and the right and the good can be congruent. Congruence is the solution to the claim of MPJ in *A Theory of Justice.*

This leads to the question of what a person's good is. To answer this question, Rawls introduces the idea of a rational plan of life into his theory. First, he contends that our life is not fragmented and disconnected. Rather, it has a unity grounded in a plan of life. As he puts it, "a person may be regarded as a human life lived according to a plan." (TJ:408/358 rev.) This plan of life provides a framework to make sense of our intentional actions, define our identity, and determine the meaning of our good. "A rational plan of life establishes the basic point of view from which all judgments of value relating to a particular person are to be made and finally rendered consistent." (TJ:409/359 rev.) A person is happy "when his plans are going well." (TJ:409/359 rev.) Moreover, "an individual says who he is by describing his purposes and causes, what he intends to do in this life." (TJ:408/358 rev.) It follows from this that most of our reasons for action can

be explained as deriving from a desire to form and to execute our plans of life successfully because "the rational plan for a person determines his good." (TJ:408/358 rev.)

The structure of a plan of life is complex.[6] Several features are particularly noteworthy. First, a plan can give a sense of unity to one's life. This can be understood in two dimensions. The first one is horizontal. We make decisions everyday. These decisions reflect our preferences for particular ends and interests, which in turn provide meaning to our life. Our ends will sometimes conflict with each other, and priority has to be established. According to what criteria do we make our choice? It depends upon our well-thought-out plans which provide a frame of reference to organize our activities and rank different desires so that our ends can be fruitfully combined into a coherent scheme of conduct. "In this way a family of interrelated desires can be satisfied in an effective and harmonious manner." (TJ:410/360 rev.) The second dimension is vertical. Our life is constituted by the past, the present and the future. The self-identity of different times is united by our plans of life. The plans render our life a kind of narrative which makes our actions intelligible. An example can illustrate this point. Suppose after careful deliberation I determine to live a philosophical life. I want to make myself a philosopher. To realize this plan, I will guide my actions with this aim in mind. I may do a degree in philosophy, read relevant books, meet friends with similar interests, or even mimic an admired philosopher's life style. My whole life will revolve around this end. Desires that are congenial to it will be

---

[6] It is also arguable if a person's life can be described as a plan. See Charles Larmore, "The Idea of a Life Plan" in *Human Flourishing* ed. E. Paul, F. Miller & J. Paul (Cambridge: Cambridge University Press, 1999), pp.96-112.

encouraged while those disturbing be weeded out. I will even identify myself with this ground project.[7] In short, these two dimensions interweave, and give my life a sense of continuity and wholeness.

Furthermore, a long-term plan normally consists of a hierarchy of specific sub-plans which are carried out at different times. However, in reality people may not be able to articulate a clear hierarchical structure for their interests and aims. Their plans may be mixed and conflicting. That explains why people feel as though they were being torn apart by different ends when they come into conflict. It is also possible that people do not have a *master* plan which encompasses all sub-plans. A life full of conflicting plans could be a terrible mess. Or even if a person has a plan, he may be unconscious of its existence and its regulative force. After all having a plan is a matter of rationalisation of our life experience. It is a complicated construction rather than a natural fact. It involves substantial work of deliberation and self-understanding.

Nevertheless, we should note that a well planned life as such is not necessarily more desirable than an unplanned one. The idea of a plan is formal and should not be assessed until its substantive content has been filled in. Moreover, as Rawls admits, "it is not inconceivable that an individual, or even a whole society, should achieve happiness moved entirely by spontaneous inclination." (TJ: 423/372 rev.) But in most cases and for most people, they see themselves as one person leading a life with a conscious plan.

---

[7] The idea of ground project can see Bernard Williams, "Persons, Character and Morality" in *Moral Luck* (Cambridge: Cambridge University Press, 1981), pp.1-19.

Finally, there is a parallel between a plan of life and a conception of the good. For Rawls, a conception of the good refers to a system which defines what is valuable in human life. It "normally consists of a more or less determinate scheme of final ends, that is, ends we want to realize for their own sake, as well as attachments to other persons and loyalties to various groups and associations." (PL:19) It informs our conduct, and gives meaning to our life as a whole. This account of a conception of the good is basically the same as a rational plan of life defined above. That is why Rawls sometimes uses them interchangeably. For instance, in explaining the subjective circumstances of justice, he says that "these plans, or conceptions of the good, lead them to have different ends and purposes, and to make conflicting claims on the natural and social resources available." (TJ:127/110 rev.)

So far we have only discussed the general structure of a plan of life. But under what conditions can a plan of life be described as rational? This question is crucial. Remember that our actions are directed by our good, and our good is in turn determined by our rational plan of life. In other words, the rationality of action depends upon the rationality of a plan of life. As Rawls expressly points out, "a person's interests and aims are rational if, and only if, they are to be encouraged and provided for by the plan that is rational for him." (TJ:409/359 rev.) But unlike Hume, Rawls does not hold that reason can only be used to evaluate the effectiveness of means to realize an end. It can be used to assess our ends as well. What Hume advocates is an *instrumental* conception of rationality, according to which an action is rational when it best satisfies the *existing* desires that we have. Reasoning may enable us to determine the most effective ways of attaining our ends, but those ends themselves are fixed by our desires. Rawls

differs from Hume by claiming that an action is rational if and only if it can best promote our good derived from our rational plans of life.[8] The existing desires do not have to be our good. For we may be *irrational* in the sense that our existing desires are contrary to our deliberate and informed plan of life. Rawls's conception is what I call *prudential* rationality.

To establish this conception of rationality, we need to know what a rational plan of life is. Rawls proposes that a person's plan of life is rational if and only if it can satisfy the following two conditions:

> (1) It is one of the plans that is consistent with the principles of rational choice when these are applied to all the relevant features of his position, and (2) it is that plan among those meeting these conditions which would be chosen by him with full deliberative rationality, that is, with full awareness of the relevant facts and after a careful consideration of the consequences. (TJ:408/358-359 rev.)

A rational plan of life hinges on principles of rational choice and the principle of deliberative rationality. The former includes the principle of effective means, of inclusiveness and of the greater likelihood of success. The latter refers to a certain attitude of deliberation which requires the agent to take carefully all the relevant facts and possibilities into account and employ the best means to realize his most important desires under favourable conditions. In this deliberative process, it is also assumed that there are no errors of calculation and the facts are correctly assessed. Apart from these principles, Rawls adds one more principle to help us determine our rational plan of life. It is "the Aristotelian Principle" which

---

[8] This distinction is borrowed from Scheffler, *Human Morality*, p.73.

states that "human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity." (TJ:426/374 rev.) In short, Rawls's aim is to use these principles of rationality to define human goods so that it can provide justification for primary goods and for our reasons for action.

It is however noteworthy that after lengthy discussion, Rawls concedes that although the rational principles can set up guidelines for reflection and narrow down the scope of rational plans, they cannot conclusively tell us how we should live. With regard to everyone's final ends, we may have to choose for ourselves from a maximal class of plans. This is because these principles are quite formal and many plans can be compatible with them. They are not substantive enough to determine what sorts of specific ends are most rational. At some point rational deliberation will reach its limit, and individuals may have to choose among a range of choices. As Rawls states, "it is clearly left to the agent himself to decide what it is that he most wants and to judge the comparative importance of his several ends." (TJ:416/366 rev.) As a consequence, different people will have different conceptions of the good. "Many things may be good for one person that would not be good for another." (TJ:448/393 rev.)

This kind of indeterminacy is the nature of prudential rationality. As our discussion proceeds, we will find that this feature of rationality has a great impact on the argument for stability. But it is understandable why Rawls holds this view. First, it matches our considered judgment that a liberal society should allow people to lead their lives in their own way. If the content of rational principles is thick to such an extent that it provides a definitive determination for our plans, it

will contradict the idea of freedom of choice. Second, it is unnecessary because conceptions of the good have no influence on the justification of principles of justice. Thus, "there is no need to set up the account of the good so as to force unanimity on all the standards of the rational choice." (TJ:447/393 rev.) Rawls further concludes that "this indeterminacy is no difficulty for justice as fairness, since the details of plans do not affect in any way what is right or just." (TJ:449/394 rev.)

## 3 Reason/Motive Internalism

If my above account is correct, it seems that Rawls intends to hold an *internalist* view concerning motivation and practical reason for action. According to Williams, internalism is the view that an agent A has a reason to do Ø if and only if A has some desires deriving from his subjective motivation set the satisfaction of which will be served by his Øing.[9] Or if we follow Darwall's classification, a reasons/motives internalism model holds that a necessary condition of *p*'s being a reason for A to do Ø is that A can have, and under suitable conditions would have, some motivation to do Ø by virtue of a suitable awareness of *p*.[10] Internalism emphasises that there is a close connection between practical reason and motivation. As Korsgaard explains, for an internalist, "if I judge that some action is right, it is implied that I have, and acknowledge, some motive or reason for performing that action. It is part of the sense of the judgment that a

---

[9] Williams, "Internal and External Reasons" in *Moral Luck*, pp.101-102.

[10] Stephen Darwall, "Reasons, Motives, and the Demands of Morality: an Introduction," in *Moral Discourse and Practice : Some Philosophical Approaches* ed. Stephen Darwall, Allan Gibbard, and Peter Railton (New York: Oxford University Press, 1997), p.307.

motive is present."[11] The reason why an action is right provides an agent with a motive for doing that action. The capability of motivation thus sets a requirement on practical reason. When the agent is not motivated by a moral judgment, there is no point for a third person to say that the agent has a reason to do that action. On an *externalist* theory, by contrast, there is no such connection between reason and motive. The reason why an action is right is separate from the motive for doing it.

Rawls does not explicitly lay out which view he holds. It is reasonable for us to believe that he is sympathetic to the internalist position though. First of all, the fact that Rawls takes moral stability so seriously itself suggests that moral justification is inseparable from the concern of motivation. As he writes, "however attractive a conception of justice might be on other grounds, it is seriously defective if the principles of moral psychology are such that it fails to engender in human beings the requisite desire to act upon it." (TJ: 455/398 rev.) We should note that how strong the desire to act justly actually is should not be a separate question in establishing a conception of justice, or to be dealt with after its establishment. Rather, the motivating force must be inherent in the conception of justice. A justifiable conception of justice must be able to provide rational agents with sufficient moral motive to act. This is indeed an internalist requirement of justification.

The second evidence comes from Rawls's account of the strains of commitment. The strains of commitment state that the parties in the original position should choose a conception of justice which they can adhere to when the

---

[11] Korsgaard, "Skepticism about Practical Reason" in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), p.315.

veil of ignorance is lifted. "They cannot enter into agreements that may have consequences they cannot accept." (TJ:176/153 rev.) This implies that the rational parties will only choose those principles which they can have sufficient motive to comply with. But whether they can have such motive depends upon the reasons provided by the principles. The source of commitment is embodied in the conception of justice. So, in comparing different conceptions, the parties will consider which conception can best promote their fundamental interests and which one is psychologically less motivating. This is actually what internalism dictates: when the parties agree to endorse a specific conception of justice, they have the accompanying motive to abide by the principles. In other words, if a conception of justice fails to meet the strains of commitment, it is inherently unjustified.

Finally, Rawls clearly objects to modern intuitionism represented by R.D.Ross who holds an externalist position. For Ross, our duty is a distinct and unanalyzable property which can only be known by rational intuition. But knowing our duty does not necessarily entail the motive for performing it. Moral knowledge is one thing and motivation another. Ross presumes that people have a distinctive sense of right to fulfil one's duty. Moreover, this motive is the highest and purest desire to respond to our duty, which is not derived from, or dependent upon any other desires. Although Ross admits that a moral motive could be triggered by what you are told is your duty, "it would be possible to have that intuition and not be motivated by it. The reason why the act is right and the motive you have for doing it are separable items."[12] Rawls argues that this is

---

[12] Korsgaard, "Skepticism about Practical Reason," p.316.

untenable because it fails to explain why rational agents must have the compelling desire to act from the sense of right if it is entirely detached from our good. For Rawls, "a perfectly just society should be part of an ideal that rational human beings could desire more than anything else once they had full knowledge and experience of what it was." (TJ:477/418 rev.) Contrary to intuitionism, a reasonable conception of justice must answer to our good so that it can motivate us to do what is just. Rawls's objection to externalism further confirms my claim that he is a reason/motive internalist.

It should be noted that the requirement of internalism has great implications for moral justification, that is, the question of justification cannot be set apart from the question of motivation. A conception of justice should be able to be justified from an impartial point of view, as well as providing an effective desire for action. But once this internalist position is combined with prudential rationality, it poses a great challenge to Rawls. Before going into this point, let me first sum up what I have said so far.

At the beginning, I state that what moral stability aims at is the motivational priority of justice, a claim about practical reason for action. This directs us to explore the content of rationality. I then argue that Rawls holds a conception of prudential rationality, according to which an action is rational if and only if it can best promote one's good defined by his rational plan of life. The internalist principle stipulates that nothing can be counted as a reason for an agent to act unless it is capable of motivating him to do so. Putting prudential rationality and internalism together, it implies that a political principle can claim to be justified if and only if the reasons for the principle can motivate rational agents to act justly,

and this in turn requires that the reasons must be in some ways related to people's rational plans of life. To explain one's practical reason for adhering to the principle of justice, we need to place his reason in the context of his plan of life. Of course, like Williams's account of a subjective motivational set, a plan of life can contain things such as projects, different sorts of commitments and personal loyalties. There is no supposition that the conception of the good must be egoistic in nature. The point is that there is an internal link between moral motive and our good.

However, we should note that the internalist requirement does not entail the overridingness of the sense of justice. It is possible that under certain circumstances the motive for a particular conception of justice may be outweighed by other considerations. The claim of MPJ is much more demanding than the internalist requirement. It demands that a conception of justice can claim to be justified only if the sense of justice could be regarded by rational agents as the most compelling motive for action stemming from their subjective motivational set. It must show that it is normally rational for the sense of justice to take priority over other competing desires.

Rawls would, however, confront a serious challenge if the above analysis stands: given that people have different conceptions of the good in a pluralistic society, how can they, as rational agents, commonly agree that insofar as they have reason to do what justice demands, moral reasons should always outweigh their rational self-interest? How can the sense of justice be always overriding if "the stability of a conception depends upon a balance of motives"? (TJ:454/398 rev.)

Faced with this challenge, the sense of justice must be shown to occupy an essential place in every agent's rational plan of life. The moral motive should not be understood as something alien to a person's conception of the good. Rather, it is constitutive of our good. Otherwise, it cannot explain why an agent, from his own point of view, should have such a compelling motive to act on what justice dictates. Therefore, to achieve stability, Rawls tells us, "what is to be established is that it is rational (as defined by the thin theory of the good) for those in a well-ordered society to affirm their *sense of justice as regulative of their plan of life*." (TJ:567/497 rev., my emphasis) Rawls puts his hope on congruence of the right and the good. When the desire to act justly is also regulative of a rational plan of life, then the sense of justice can be regarded as the most important good in one's conception of the good. In this case, rational people can surely have the strongest desire to act justly.

To what extent Rawls's strategy succeeds need not concern us here.[13] The foregoing discussion, however, immediately gives rise to a more urgent question, that is, why is the claim of MPJ so essential to justification? Does Rawls set himself a daunting task which does no good to his whole project? After all, few critics share with Rawls the view that moral stability should play such a fundamental role in political philosophy. To dispel these doubts, we need to explore further the relations between justification and motivation.

## 4 The Normative Question and Justification

In this section, I shall attempt to offer some arguments to explain why MPJ is

---

[13] This will be examined in Chapter 4.

integral to political justification. First of all, it is related to Rawls's understanding of the nature of a theory of justice. According to Rawls, the first and foremost task of political philosophy is to construct a conception of justice to regulate our social cooperation. This conception defines our basic rights and duties, and determines the proper distribution of the benefits and burdens of cooperation. They are the most fundamental and ultimate principles for the social basic structure. Structurally these principles are expected to have lexical priority over other interests.

Furthermore, these political principles are normative. They make claims on us. They are practical and action-guiding. They prescribe what we ought or ought not to do. They command our allegiance. Sometimes moral commands are stringent and demanding to such an extent that we have to adjust, or even sacrifice our preferences, personal interests and important projects. We also have obligations to respect our fellow citizens' rights and liberties, pay tax and fulfil different kinds of duties, and treat other people as equals. In short, the political institutions provide a comprehensive framework for our political world and determine the legitimacy of our social actions.

More importantly, unlike other moral norms, the application of political principles inevitably involves the exercise of coercive power. The principles are embodied in the political and legal system. In a closed and self-contained society, citizens have to obey the rules. This does not mean that we do not have freedom of action. Yet even our right to liberties are protected by coercive force. If someone violates other citizens' rights, he will be punished. Undoubtedly, other moral norms have normative force as well. When we say, for instance, lying is

morally wrong, it implies that everyone should not lie. When someone does lie, in most cases what he suffers is self-blame or public disapproval from others, but not legal punishment.

Finally, as Rawls suggests, the basic structure of society has deep influence on our life prospects and in fundamental ways shapes our character and way of life from the very beginning. It will affect the future of subsequent generations as well.

Since living in a political community and adhering to its principles place such powerful normative constraint on us, we must ask why political principles should have such claim on us as rational and free agents. The rationality involved here refers to prudential rationality defined earlier. A free agent, according to Rawls, has two salient features. First, the agent regards himself as a self-originating source of valid claims. It means that people believe their claims "carry weight on their own without being derived from prior duties or obligation owed to society or to other persons, or, finally, as derived from, or assigned to, their particular role."[14] They are independent and autonomous. Their wills are not dependent upon anyone else's, and obligations are self-imposed. As equal moral persons in terms of their moral capacity, they consider themselves as having an equal right to make claims on the design of social institutions. Second, free agents are viewed as having the moral power to form a conception of the good. They have the capacity to form, revise, and rationally pursue a plan of life. They can stand apart from existing desires and critically assess their ends. This is close to

---

[14] Rawls, "Kantian Constructivism in Moral Theory," in *Collected Papers*, ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.330.

the idea of personal autonomy. A free person is the one that has second-order desires regarding the ordering of his first-order desires.[15]

Now suppose that I am a rational and free person and fully understand the nature of political principles, I put the following question to myself: is it always rational for *me* to act on the requirement of justice even if it is inconsistent with the dictates of my rational good defined by my plan of life? Why must I give priority to justice over my ends and interests? This is what Korsgaard calls *the normative question*. It asks where the sources of moral claims come from, and how they can have such normative force to outweigh an agent's other desires. The normative question is the question of justification. It purports to probe the ground of morality and ask why it can make claims on us. It is worth noting that the question must be asked from an agent's first-person perspective. As Korsgaard explains,

> The normative question is a first-person question that arises for the moral agent who must actually do what morality says...The answer must actually succeed in addressing someone in that position. It must not merely specify what we might say, in the third person, about an agent who challenges or ignores the existence of moral claims.[16]

The reasons provided for action must be endorsed from the rational agent's own point of view. A political theory should not only be able to *explain*, from a third-person point of view, why an agent has reasons to do what morality dictates,

---

[15] Further discussion on the second-order desire can see Harry Frankfurt, "Freedom of the Will and the Concept of a Person" in *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), pp.11-25.

[16] Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), p.16.

but also be able to *convince* the agent why it is rational to care about morality, and care about it so imperatively. What justification aims to achieve is to convince someone in a particular moral position. We want the agent himself to see the normative force of a moral argument. Only then can the agent generate sufficient motive to comply with the political principles. I think this is another way to reiterate the idea of internalism, which stresses that reasons for action must address agents from their own point of view. This does not mean that the justificatory reasons must be partial or agent-relative. Deliberating from the first-person perspective sets no prior limit to what kinds of reason should be included in justification. For a rational moral agent, he can in principle make use of any reasons available to justify a political principle, including impartial reasons.

It should be noted that the normative question is concerned with the rationality of actions, not with the actual actions of every agent. No theory can claim to be justified if the necessary condition of justification is that its moral commands must be *actually* accepted by every agent from his own point of view. This is implausible. For different people act differently for different reasons. No one or institution could guarantee that every agent would always act for the same reason. People can act irrationally. So, the requirement must be that a theory can be justified only if it can motivate every *rational* agent to act from his own perspective. Furthermore, it is also undesirable that the answer to the normative question should be ultimately dependent upon the agents' actual psychological motivational systems because it would then dilute the importance of the normative question and would trivialise the real tension between morality and self-interest. For according to this view, what I am morally required to do *must* coincide with

what I actually desire to do. No conflict exists between the two sides. But this is simply untrue. What one is obligated to do can have deep conflict with what he currently desires to do. It is exactly because of this possible conflict that the problem of stability becomes a salient issue.

It is then clear that the question of motivation is inseparable from the question of justification. If there are moral grounds for a conception of justice, then it must be rational for us to be motivated to act on those grounds. The motivational force of a conception constitutively determines its desirability. A justifiable theory must explain why someone, as a rational and free agent with his own rational plan of life, should have reasons to take seriously the standpoint of justice. Failing to offer a satisfactory account, justification could not get off the ground and the binding force of justice would be weak and unstable. As Scanlon rightly suggests, an adequate moral philosophy should not stop at assuming that morality is nothing more than a preference people happen to have. "It must make it understandable why moral reasons are ones that people can take seriously, and why they strike those who are moved by them as reasons of a special stringency and inescapability."[17] This is exactly what moral stability dictates, that is, a stable conception of justice must be able to generate in each rational agent an effective sense of justice to outweigh other desires. The overridingness of the sense of justice is an inherent requirement of justification. In this sense, as Dworkin rightly says, "the search for the foundation of a political theory, in the sense I have in mind, is sometimes described as the problem of finding *motivation* for the

---

[17] Scanlon, "Contractualism and Utilitarianism," in *Utilitarianism and Beyond* ed. Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982), p.106.

theory."[18]

## 5 The Status of Impartiality

The claim of MPJ in political justification has been seriously challenged. It is argued that it is a mistake to confuse the question of motivation with that of justification. The task of justification is to pursue truth or search for impartial reasons that can be shared by everyone. It is indifferent to the question of whether and to what extent it can have the corresponding motivational force to command people's allegiance. Motivational efficacy should have no impact on the justifiability of a conception of justice. According to Mendus,[19] Barry holds this view because he claims that:

> My concern is with truth, not with popularity. If I am right, justice calls for radical change...But how strong the desire to behave justly *actually* is, when it comes into competition with other desires, I leave open. I claim only to tell you what justice is; what you do about it, if you believe me, is up to you.[20]

In another occasion where Barry responds to his critics, he again states that "if somebody is totally unmoved by the elementary thought that I have expressed, then of course the theory will not speak to him. But the theory can explain why it is justifiable to do whatever is necessary to restrain such people."[21] Does Barry pose a valid critique of Rawls? Below I will make a comparison between Rawls

---

[18] Dworkin, "Foundations of Liberal Equality," in *The Tanner Lecture on Human Values Vol. XI* (Salt Lake City: University of Utah Press, 1990), p.5.

[19] Mendus, *Impartiality in Moral and Political Philosophy* (New York: Oxford University Press, 2002), p.10.

[20] Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), p.115, my emphasis.

[21] Barry, "Something in the Disputation not Unpleasant " in *Impartiality, Neutrality and Justice* ed. Paul Kelly (Edinburgh: Edinburgh University Press, 1998), p.237.

and Barry, and argue for the necessity of MPJ in justification.

To begin with, I must say Mendus's observation is a bit inaccurate. What Barry exactly says is that people's *existing* motivation has no necessary bearing on the justifiability of a conception of justice. For Barry, the proper task of political philosophy is to quest for right political ideals. A political theory has no duty to warrant that every citizen must be *actually* motivated to comply with the principles. This is because there are at least two possible reasons to explain why a person lacks incentive to act justly. The first is that the person is self-interested, irrational, or unreasonable. The second is that the political principle itself is unjustified or unreasonably demanding. In the second case, the theory at stake has reason to step back to re-examine its arguments. But if we have good reasons, as Barry claims, to believe that our theory is true or right, there is no point in making concessions simply because of the reasons of the first type of case. And as a matter of fact, people do act irrationally or unreasonably. As Korsgaard rightly points out, even in a standard account of instrumental rationality, an agent might choose means insufficient to his given end even though he knowingly understand the relevant causal relations in the case. The agent may simply fail to transmit the motive force from the operation of means-end reasoning. This is because:

> The necessity, or the compellingness, of rational considerations lies in those considerations themselves, not in us: that is, we will not necessarily be motivated by them...So a person may be irrational, not merely by failing to observe rational connections—say, failing to see that the sufficient means are at hand—but also by being "wilfully" blind to them, or even by

being indifferent to them when they are pointed out.[22]

Once we allow the case of true irrationality, it is unreasonable to expect a political theory to be actually able to motivate all people to do what justice requires. A correct formulation should hold rather that a conception of justice is justified only if it can effectively motivate *rational* people to act in accordance with the principles of justice. Barry accepts this motivational requirement indeed. He makes it clear that any theory of justice must presuppose an account of the motive for behaving justly. The stipulated account will in turn substantially determine the form and content of the principles of justice. As he remarks, "because of the practical nature of justice, a theory of the motivation for being just must at the same time be a theory of what justice is. For the content of justice has to be such that people will have a reason for being just."[23] Take justice as mutual advantage as an example. According to this theory, the major motive for agents to comply with the rules of justice is derived from their belief that doing so is in the long term a more effective way to promote their conceptions of the good. Based on this motive, a set of rules is regarded as just if general compliance with the rules would be more advantageous to every participant than other alternatives.[24] Justice as impartiality has another story of motivation. It has a different starting point. People are not presumed to be moved simply by their wish to advance their conception of the good. Rather, they are motivated by the desire for reasonable agreement. Following Scanlon's path, Barry supposes that it is a widespread fact that agents have a desire to act according to rules that could not reasonably be

---

[22] Korsgaard, "Skepticism about Practical Reason," p.320.

[23] Barry, *Theories of Justice* (California: University of California Press, 1989), p.359.

[24] Barry, *Justice as Impartiality*, p.46.

rejected by others similarly motivated. They will therefore only accept those rules that "no one could reasonably reject as a basis for informed, unforced, general agreement" as just. Without this moral motive, justice as impartiality cannot get off the ground in the first place.

Whether Barry's argument for justice as impartiality is sound is not my concern here. My point is that Barry, like Rawls, takes the problem of motivation seriously and views it as inseparable from the question of justification[25] Notwithstanding this similarity, there are two important differences between Barry and Rawls that is noteworthy. First, Barry holds that the impartial motive to reach reasonable agreement is a distinct and independent desire which has no connection with our conception of the good. It reflects our moral commitment to treat everyone as equals. "The motive is the desire to act justly: the wish to conduct oneself in ways that are capable of being defended impartially."[26] Barry, however, argues against the view that the source of moral motive must in some way be related to one's conception of the good. For Barry, recognising something to be just should itself be sufficient to motivate an agent to comply with the just rules. It is unnecessary and undesirable to explain one's acting justly in terms of the interest of the agent. The impartial motive, by definition, is contrary to the partial concern of one's conception of the good. People can simply act out of a sense of justice. If the motivation for being just is derived from its long-term advantageousness to the self-interest of the agent, then it inevitably falls into a variant of justice as mutual advantage which is, for Barry, indefensible.

---

[25] Barry makes this point clear in *Theories of Justice*, pp.357-66.

[26] Barry, *Theories of Justice*, p.363.

Rawls is at this point sharply different from Barry. For Rawls, one's reason to act justly cannot be entirely detached from one's conception of the good. As shown above, Rawls holds that our reason for action is dependent upon the good defined by our rational plan of life. Therefore, if the sense of justice cannot be properly integrated into the agent's conception of the good, it is hard to explain why the agent must act in accordance with principles of justice.[27] That is why Rawls claims that the "doctrine of the purely conscientious act is irrational." (TJ:477/418 rev.) It is irrational because this doctrine presumes that the moral motive, as the highest motive, requires us to do "what is right and just simply because it is right and just." It is stipulated not to have any relation with our plans of life. Rawls believes that this view fails to account for our moral and psychological experience. It is natural and reasonable for us to have an interest in advancing our conception of the good. If the moral motive is so distinct and detached from our pursuit of well-being, why should we, as rational agents, take it so seriously? It seems that "the sense of right lacks any apparent reason; it resembles a preference for tea rather than coffee." (TJ:478/418 rev.) [28] Nevertheless, this does not mean that Rawls endorses a Hobbesian conception of justice as mutual advantage. Rawls is an impartialist in the sense that the principles of justice must be agreed by free and equal beings on an equal footing in the original position. He also stresses that people are motivated to act by the sense of justice specified by the conception of justice in a well-ordered society. What Rawls insists is that the effectiveness of the sense of justice should be

---

[27] Influenced by Scanlon, the later Rawls seems to have made some change when he talks about the existence of principle-dependent desires and conception-dependent desires. See PL:82-84.

[28] For Barry's defence for the doctrine held by Ross and Prichard, see "John Rawls and the Search for Stability", p.884.

explained in terms of its regulative role in people's rational plans of life. Thus, "in the light of the theory of justice we understand how the moral sentiment can be regulative in our life." (TJ: 478/418 rev.)

Comparing these two accounts of moral motivation, I intend to think that Barry's view is more problematic. While Barry agrees that "our account of the nature of justice cannot be separated from the question of motivation," he stresses that the motivation for acting impartially is itself sufficient to explain why it is rational for agents to conform their conduct to the demands of justice. For the desire to reach informed and reasonable agreement without appealing to personal advantage is actually widely shared. Barry therefore concludes that "what I am saying is that the desire to be able to justify our conduct in an impartial way is an original principle in human nature and one that develops under the normal conditions of human life."[29] Barry simply supposes that most of us are impartialists.

But this position immediately gives rise to some difficulties. First of all, how would Barry respond to those who do not share his assumption of moral motive? An impartial moral motive is not simply a description of psychological fact. It reflects, as Barry stresses, a fundamental commitment to the equality of all human beings.[30] We grant all agents equal right to choose principles of justice which cannot be reasonably rejected because we have already accepted the notion of equal worth of human beings. Justice as impartiality commands us to transcend our differences in social background and natural endowments, and to treat one

---

[29] Barry, *Theories of Justice*, p.364.

[30] Barry, *Justice as Impartiality*, p.8.

another as morally equal. The reason for an agent to endorse the criterion of reasonable acceptability must be that he has already made a moral commitment that he should not enjoy any special privilege in determining what justice is. As Barry remarks, "only on this basis can we defend the claim that the interests and viewpoints of everybody concerned must be accommodated."[31] Thus, the moral motive itself embodies a commitment to equality. This commitment, however, calls for justification. We need to know the basis of equality, and why it is rational for people with different bargaining power and conceptions of the good to adopt this impartial perspective. Apparently, the answer to this question is crucial to Barry's project. As Mendus succinctly summarizes, the task for political impartialism is:

> To show why those who are not themselves impartialist might nonetheless accept an impartialist political order, why they might accept it as genuinely just, and why they might concede that its demands take priority over the conflicting values endorsed by their own comprehensive conceptions of the good.[32]

What surprises us is that Barry provides very little discussion on this fundamental issue. More accurately speaking, he does not bother with it. He simply presumes that most people are generally motivated to act from an impartial motive. As he says, people "are actuated solely by a motive that has force with almost all of us to some degree, the desire to act in ways that can be defended to others."[33] But how could that be? If it is an empirical claim, then it is far from

---

[31] Barry, *Justice as Impartiality*, p.8.

[32] Mendus, *Impartiality in Moral and Political Philosophy*, p.3.

[33] Barry, *Justice as Impartiality*, p.10.

clear how many people actually accept this particular conception of impartialism. Countless counter examples suggest that in advanced capitalist societies many people do not share Barry's impartialist sentiment. Above all, no matter how prevalent a motive actually is, we still need a justification. In particular, we need an answer to the normative question as to why a rational agent has reason to be a liberal egalitarian. Unfortunately, Barry simply ignores this question.

Barry then could not dismiss those non-liberal egalitarians as unreasonable because this would put his own proposition into question. For the content of reasonableness is essentially characterized by a commitment to equality. Nor can he say that the non-egalitarians are necessarily sexist, racist or egoist. For this need not be the case. After all the real challenge of justice as impartiality comes from other moral theories. Take Rawls as an example. Rawls's justice as fairness is regarded by Barry as the best-known, the most influential and the most fully developed variant of justice as impartiality. Barry basically concurs with Rawls that the distribution of natural talents and social background are arbitrary from a moral point of view, and these differences should not be counted in determining principles of justice. This expresses the ideal of moral equality. Yet in the past three decades of dispute about Rawls's theory of justice, one of the most controversial issues is about the desirability of this characterisation of egalitarian commitment. It has been seriously questioned by libertarianism, communitarianism, and Marxism among others.[34]

---

[34] For example, Nozick, *Anarchy, State and Utopia* (New York: Basic Books, 1974); Michael Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982); Kai Nielsen, *Equality and Liberty* (Totowa: Rowman & Allenheld, 1985).

If Barry wants to defend his liberal impartialism, he must offer substantive reasons to account for the moral basis of equality and explain why rational people can have sufficient motive to commit to liberal impartiality. It is not enough to just *presuppose* that we have such a commitment and the corresponding motive. Barry's strategy to disconnect the impartial motive from people's conception of the good makes this task particularly difficult. The reason is that if impartial motivation has no positive connection with people's plans of life, it is hard to explain why rational people should be convinced of justice as impartiality, especially when the moral demand is in deep conflict with people's conceptions of the good. Therefore, when Barry distances himself from justice as mutual advantage by emphasizing the independence of the moral motive, he also cuts the sense of justice off from our conceptions of the good. When being challenged by non-impartialists, Barry may say: "you either take it or leave it." But this is exactly what Rawls's dissatisfaction with the doctrine of the purely conscientious act expresses, namely that "the sense of right lacks any apparent reason." (TJ:478/418 rev.)

Barry is aware of this problem. He thus develops another way to defend himself by saying that:

> The basis of justice is institutional, I have argued, and institutions normally deploy sanctions to provide an additional motive for compliance. It is not, therefore, necessary that everyone should be moved by a sense of justice so long as the gap can be filled by deliberately created incentives for compliance.[35]

---

[35] Barry, *Theories of Justice*, p.366.

But this response is of little help. For if we accept this solution, it is equivalent to conceding that justice as impartiality cannot be justified to those who lack the impartial motive in the first place. Barry's recourse to sanction as an additional motive for compliance may to some extent resolve the problem of social stability at the practical level, but it does nothing to settle the problem of moral stability at the justificatory level. It can only apply to those who have already accepted the liberal conception of equality. This argument is too limited to justify impartialism to free and rational agents.

Now, for the sake of argument, suppose Barry is right that many people do have a desire to act impartially, the story still does not end. For he needs to show how this moral motive can be overriding. That a person has a reason to commit to impartiality does not mean that it can necessarily take priority over other desires. The impartial motive is only one of the motives in people's subjective motivational set. Granted that we accept the primacy of justice as a theoretical requirement, the motivational priority of justice is a substantive one. There is no guarantee that simply in virtue of the role of justice, a conception of justice necessarily takes precedence over other conflicting interests. This brings us back to the primary question of moral stability. This is the claim that "the stability of a conception of justice depends upon a balance of motives: the sense of justice that it cultivates and the aims that it encourages must normally win out against propensities toward injustice." (TJ:454/398 rev.)

We see that Rawls acknowledges the tension of different motives and seeks a way to resolve it. This is not a practical issue about how to ensure compliance after the political principles have been justified. Instead, the question at issue is

about justification itself. If the justification of justice, as Barry himself recognizes, is inseparable from the problem of motivation, then the priority of moral motive is a prerequisite for the priority of justice as impartiality. Unfortunately, Barry does not offer any substantive argument for this crucial issue. He simply takes it for granted:

> Self-interest cannot be expected to bring about just institutions in general, so it is crucial that the sense of justice should operate there [justification of principles of justice]. Fortunately, all that is often necessary is that those whose own interests are not directly affected should support the course of impartial justice. [36]

However, this account is truly puzzling. On the one hand, Barry agrees that the sense of justice must be the regulative motive in justification; on the other hand, he claims that for this purpose it is enough if those whose interests have no direct conflict with impartial justice will endorse the priority of the sense of justice. But this response does not answer the question at all. The problem of motivational priority becomes an issue because there is a conflict between moral motive and self-regarding desires. As Mendus rightly describes the problem, the real question posed to moral impartialism is:

> Expressed in the agent's self-directed question 'why should I act on the motivation to do what impartial morality dictates rather than on the motivation to act partially?' Since this question arises even (indeed especially) for those who accept the importance of impartial demands, it forces us to consider the source and extent of impartialism's motivational power.[37]

---

[36] Barry, *Theories of Justice*, p.366.

[37] Mendus, *Impartiality in Moral and Political Philosophy*, p.3.

The implication of the preceding analysis is that the problem of moral stability is not a peculiar issue only inherent in Rawls's theory. It is a normative question that every political theory needs to answer. Barry, contrary to Mendus's interpretation, does acknowledge the importance of this question. Nevertheless, his solution fails because he simply presupposes the priority of impartial motives rather than justifies it. Rawls's solution in *A Theory of Justice*, on the contrary, is to attempt to link up the moral motive with the conception of the good. To justify the motivational priority of justice as fairness, Rawls believes that the sense of justice must be defended as a regulative good in every rational agent's rational plan of life. As he expressly states,

> If within the thin theory it turns out that having a sense of justice is indeed a good, then a well-ordered society is as stable as one can hope for. Not only does it generate its own supportive moral attitudes, but *these attitudes are desirable from the standpoint of rational persons who have them when they assess their situation independently from the constraints of justice*. This match between justice and goodness I refer to as congruence. (398-399/350 rev., my emphasis)

For Rawls, the source of moral motive is not something alienated from our good. Rather, it is desirable because it can be understood as a higher-order regulative good from our first-person rational standpoint. This congruence approach matches very well with the idea of prudential rationality, that is, rational persons have sufficient reasons to act in accordance with the principles of justice because the action itself is the most effective way to promote our fundamental interest.

If we accept the congruence argument as essential to Rawls's project, it

creates a strong tension with a powerful interpretation of Rawls's theory as a form of Kantian deontological theory which holds that the right is justified in a way that does not depend on any particular vision of the good. The primacy of justice entails liberal neutrality among conceptions of the good. This view has been most famously presented and strongly attacked by Sandel in his seminal work *Liberalism and the Limits of Justice*. The tension is that if congruence is a necessary condition for moral stability, and moral stability is in turn a necessary condition for the justification of a conception of justice, then the motivational priority of the right is inseparable from the good. But for Sandel, this is what deontological liberalism squarely opposes. In the following, I will show that Sandel's critique is flawed.

## 6 Deontological Liberalism and Stability

Let me first explain Sandel's main thesis. Sandel holds that Rawls's theory is a type of deontological liberalism. Its core thesis is the primacy of justice, which can be understood in two different but related ways. In its moral sense, the demands of justice outweigh other values. Justice is perceived as the first virtue of social institutions. According to this view, "justice is not merely one value among others, to be weighted and considered as the occasion arises, but the highest of all social virtues, the one that must be met before others can make their claims."[38] The claim of justice is overriding. In Rawls's context, it means that the two principles of justice have absolute priority over our aims, interests, and conceptions of the good in case of conflict.

---

[38] Sandel, *Liberalism and the Limits of Justice*, p.2.

How can the moral priority of justice be justified? If the foundation of justice depends on any particular vision of the good, the proposition that the right has priority over the good will hardly be solid. Sandel thus further asserts that the primacy of justice implicates a *foundational* sense which describes "a form of justification in which first principles are derived in a way that does not presuppose any final human purposes or ends, nor any determinate conception of the human good."[39] Justification of justice must be neutral among conceptions of the good.[40] In Sandel's view, it is this second-order, foundational sense of primacy that distinguishes Rawls's deontological liberalism from other theories of justice, which in turn leads Rawls to endorse an unencumbered conception of the self. Sandel's whole critique of Rawls is based on the implausibility and undesirability of this radically unsituated self.

If we follow this interpretation, how would deontological liberalism account for the importance of moral stability and Rawls's congruence argument? If the grounds of justice are completely detached from human good, where does the moral motive originate from when people go back to their real lives? Our discussion above has clearly shown that Rawls takes these questions seriously and proposes the congruence argument to deal with them. For Rawls, to justify the motivational priority of the sense of justice, the right must be in some way related to our good. The motive to act justly is not something independent of our subjective motivational set. Otherwise, agents will lack incentive to endorse the

---

[39] Sandel, *Liberalism and the Limits of Justice*, p.3.

[40] Rawls himself has never used the term "neutrality" in *TJ*. But in *PL*, he discusses different ways of neutrality and says that his political conception of justice is a form of neutrality of aim. See *PL*, pp.190-94 for detailed discussion.

priority of justice.

Sandel owes us an account how his interpretation can make sense of Rawls's unfailing pursuit of moral stability throughout his philosophical life. Quite surprisingly, Sandel has entirely set aside this important question from his discussion. He pays almost no attention to this problem in his book-length critique of Rawls in spite of Rawls's explicit claim that the whole discussion of Part Three of *A Theory of Justice* is "to prepare the way to settle the question of stability and congruence." (TJ:395/347 rev.) There are even no entries for "stability" and "congruence" in the index of his work while Sandel quotes heavily from Part Three to justify his critique of Rawls. A reader of Sandel who has never read *A Theory of Justice* may mistakenly think that the problem of stability is not an issue at all in Rawls's theory.

There are several possible explanations for this peculiarity. Sandel may argue that the problem of stability has no importance in justice as fairness. This is simply wrong, though, because Rawls makes it clear that stability is a desirable feature of moral conception. "However attractive a conception of justice might be on other grounds, it is seriously defective if the principles of moral psychology are such that it fails to engender in human beings the requisite desire to act upon it." (TJ: 455/398 rev.) As will be expounded in the next chapter, the concern for stability in fact plays an important role in deriving the two principles of justice in the original position. Sandel has no excuse for overlooking this important issue.

Sandel's second reply might be that the moral primacy of justice has already entailed the motivational priority of justice. If the principles of justice are fully

justified in the original position, we can then say that it is most reasonable and rational for us to act in accordance with them. Moral priority implies motivational priority. Justice must trump other values should conflict arise. Therefore, certain people's lack of effective sense of justice would not affect the integrity of justice. The problem of stability then has no independent moral status.

This explanation is untenable. It is at least not Rawls's own view. First of all, this account faces a similar problem to Barry's. The original position undoubtedly represents an impartial perspective. It models a conception of free and equal moral persons in the hypothetical contractual situation. But when the veil is lifted and people are back to reality, the question of motivational priority will arise. The claim of the primacy of justice does not entail the priority of moral motivation. For the former is a formal requirement about the role of justice in social cooperation, while the latter requires a substantive account of how the sense of justice is related to our good.

Sandel might retort that as the problem of stability arises only after the veil of ignorance is lifted, it has no place in justification because the principles have already been fully justified in the original position. But this is not Rawls's view. The later Rawls particularly stresses that there are two stages in the justification of justice as fairness. Though stability arises only at the second stage, "the argument for the principles of justice is not complete until the principles selected in first part are shown in the second part to be sufficiently stable." (PL:141, footnote 7) Justification must take the second stage into account.[41]

---

[41] In the next chapter, I will argue that this two-stage justificatory structure has already existed in *A Theory of Justice*.

I believe that the only possible explanation for Sandel's overlooking of the stability problem is that it would pose a challenge to his interpretation of Rawls as a deontological liberal holding an implausible account of an unencumbered self. To make his charge coherent and consistent, Sandel cannot but argue that the justification of justice as fairness has no connection to the human good and the priority of motivation. Not doing so would seriously weaken his central thesis. My analysis, however, shows that this is not a sympathetic and faithful interpretation of Rawls.

My response to Sandel would, however, create some internal problems. If Rawls's justification has a two-stage structure, what is the proper relation between these two stages? If the first stage justification in the original position has excluded any knowledge about people's particular conception of the good, how can congruence succeed in the second stage? These questions must be answered. And when they are, we will have a different picture of justice as fairness. I will assume this task in the next chapter.

# CHAPTER 3

# THE PLACE OF STABILITY

In the preceding chapter, I have argued that the idea of moral stability, being characterized as the pursuit of the motivational priority of justice, is essential to justification. This chapter will take one more step in investigating the place of stability in justice as fairness. I will focus particularly on the role of stability played respectively in the two stages of justification.

To anticipate my argument, I am going to make three major claims. First, I argue that stability is one of the main grounds for the principles of justice in the first stage. The concern of moral motivation has direct bearing on the choice of principles in the original position. It is misleading for Rawls to say that stability is only a matter concerning the feasibility of the second stage. Second, I contend that the real force moving the parties in the original position to adopt the maximin rule actually results from moral considerations. The deliberation of parties in the original position must be guided by moral reasons if we expect them to adopt the maximin rule. The idea of grounding justification on the rational choice of self-interested persons therefore cannot account for the desirability of Rawls's principles. Finally, I argue that the second stage is also indispensable to the justification of Rawls's principles of justice. Nevertheless, the argument in this stage is not to confirm the feasibility of principles of justice derived from the first stage as Rawls suggests, but to justify the overridingness of moral motivation over other desires which cannot be fully answered in the first stage.

With these arguments, I further affirm my claim that moral stability occupies

a central place in both stages of justice as fairness. The concern for motivational priority is integral to moral justification. The structure of this chapter is as follows. In Section 1, I will present an account of the place of stability in Rawls's theory, and then refute a representative interpretation proposed by Freeman. This sets the background for the development of my own account. Section 2 will examine Rawls's claim that stability is limited to the second stage because of the concern about envy and special psychologies. I will show that Rawls's true intention in doing so is to produce an ideal environment for rational choice so that his principles can be presented as a result of consent by following the maximin rule. The maximin rule is the kernel of Rawls's contract theory. This leads us to conduct a more thorough examination of the nature of contractarianism in Section 3, where I argue that Rawls's theory involves two models of justification, namely the reasonable model and the rational model. But since Rawls holds that the reasonable is prior to the rational, justice as fairness can hardly be described as a contract theory unless the maximin rule itself is an object of rational choice by mutually disinterested persons. However, there are no decisive arguments for the parties to prefer the maximin rule to the principle of insufficient reason. The justification for Rawls's principles is then open to doubt. To avoid this predicament, I suggest that we should endorse a moral argument for the maximin rule. The parties are moved by moral reasons rather than egoistic ones to favour the conservative decision rule. Sections 4 and 5 are devoted to validating this rather unusual claim. Section 4 focuses on the lexical priority of basic liberties while Section 5 concentrates on the strains of commitment and stability. Against this background I can finally confirm that stability has already played an important role in the first stage in justifying the principles of justice. Once the first

stage argument is settled, Section 6 turns to discuss the second stage, in which I explain why the second stage of stability is integral to justification. The last section is a conclusion.

## 1 Two-Stage Justification

This section will first present Rawls's account of the place of stability in his theory, and then refute a popular interpretation suggested by Freeman. The discussion will set the background for our pursuit of an alternative interpretation of the status of stability.

In section 76 of *A Theory of Justice* entitled "the Problem of Relative Stability", Rawls draws an important conclusion after a lengthy comparison between justice as fairness and utilitarianism as follows:

> These remarks are not intended as *justifying reasons* for the contract view. The main grounds for the principles of justice have already been presented. At this point we are simply *checking* whether the conception already adopted is a *feasible* one and not so unstable that some other choice might be better. (TJ: 504/441 rev., my emphasis)

Furthermore, in the last section of *A Theory of Justice*, Rawls explains that his theory consists of three parts, each of which is intended to fulfil a specific purpose. The first part presents a theoretical structure of the original position from which a conception of justice for the basic structure is derived. Rawls's principles of justice are chosen unanimously by rational parties there. The second part discusses what sorts of social and political institutions should be established to fulfil the requirements of justice, and in what way they match our considered

judgments better than other rival theories. Finally, the third part is set to check whether justice as fairness is a feasible conception. Rawls says:

> This forced us to raise the question of stability and whether the right and the good as defined are congruent. *These considerations do not determine the initial acknowledgement of principles in the first part of the argument, but confirm* it." (TJ:580/508 rev., my emphasis)

The quotation above gives us an impression that stability is solely concerned with the feasibility of a conception of justice that has already been independently justified on other grounds. The function of stability is to confirm rather than to justify a conception of justice. In light of this view, the consideration of stability has no relevance to the justification of principles of justice. This impression is further strengthened when the later Rawls explains that his theory is divided into two stages, and stability is an issue that would only be taken up in the second stage. As he states:

> Justice as fairness is best presented in two stages. In the first stage it is worked out as a freestanding political (but of course moral) conception for the basic structure of society. Only with this done and its content—its principles of justice and ideals—provisionally on hand do we take up, in the second stage, the problem whether justice as fairness is sufficiently stable. (PL:140-41)

This two-stage structure is not a novel idea. It has been repeatedly mentioned throughout Rawls's works.[1] The first stage is responsible for working out the

---

[1] For example, see TJ:144/124 rev., 504/441 rev., 530-31/465 rev.; also see Rawls, *Justice as Fairness: a Restatement* (Cambridge, Mass: Harvard University Press, 2001), pp.88, 181.

conception of justice while the second stage deals with the problem of stability as feasibility. As justification is the first task of political philosophy, stability is naturally a secondary issue no matter how important it might be in other respects. If we accept this widely held interpretation, explaining why stability is essential to justification would be highly problematic.[2]

But why should stability be limited to the second stage? Freeman provides a ready explanation. He suggests that if social stability is introduced into the first stage and we view it as the primary subject of justice, justice as fairness would risk becoming a Hobbesian conception of justice as mutual advantage. This needs explanation. Freeman believes that concern for social stability is a common feature of the contractarian tradition. But Hobbes is different from Locke, Rousseau and Kant in the sense that he conceives of social stability as the primary subject of justice. "A just society for Hobbes is nearly identifiable with a stable social order."[3] Thus, the content of justice is defined by those norms and institutions that can most effectively achieve a stable social order. Moreover, these norms are achieved as the result of a practical compromise among essentially self-interested persons. Rawls, following the track of Kant, holds a different view. He does not view stability as the first goal of political justice. For the concern of social order alone is insufficient to account for a reasonable moral point of view. As Freeman puts it, "by itself a stable social order, however rational it may be, can be of little moral consequence if it does not rectify but only perpetuates gross

---

[2] I have examined this issue in Chapter One.

[3] Samuel Freeman, "Congruence and the Good of Justice," in *The Cambridge Companion to Rawls* edited by Samuel Freeman (Cambridge: Cambridge University Press, 2003), p.278.

injustice."[4] Introducing stability to the first stage would compromise the integrity of moral principles. For social stability is not a moral consideration. Freeman thus concludes that "a conception of justice should be worked out beforehand by relying on *independent moral considerations*. Then the question of its stability is raised to test the *feasibility* of a just society conceived along the lines of this conception."[5] This account seems to perfectly explain why stability should be limited to the second stage.

I believe Freeman's explanation is flawed. First, Freeman's account is based on a wrong assumption that stability only refers to social stability. He fails to notice that there is a distinction between social stability and moral stability, and it is the latter that concerns Rawls. Once we define *the priority of moral motivation* as the first task of stability, Freeman's worry can be settled. For Rawls's conception of stability is itself a serious moral concern. It is an inherent moral requirement for a conception of justice to demonstrate its ability to motivate people to act on justice. It is not a purely practical matter at all. Thus there is no principled reason why stability must be excluded from justification.

Secondly, Freeman's account cannot sufficiently explain Rawls's claim that stability is fundamental to political philosophy. If it is merely a practical issue concerning how to enforce principles in society effectively, there is no need for Rawls to make a philosophical turn to political liberalism. He can simply search for some more effective means of persuasion or enforcement to realize the principles independently justified. An unsatisfactory argument for stability in the

---

[4] Freeman, "Congruence and the Good of Justice," p.278.

[5] Freeman, "Congruence and the Good of Justice," p.279, my emphasis.

second stage can hardly be the main cause forcing Rawls to reformulate his philosophical argument of the first stage. Freeman is aware of this explanatory difficulty. He clarifies that what Rawls pursues is stability for the right reasons, not stability *per se*.[6] As Rawls himself expounds, "finding a stable conception is not simply a matter of avoiding futility. Rather, what counts is the kind of stability, the nature of the forces that secure it." (PL:142) Rawls is only interested in a particular kind of stability which must result from citizens' effective sense of justice. But if so, it confirms my claim that Rawls's conception of stability as a moral ideal will not compromise the integrity of political justice even if it is introduced into the first stage.

Lastly, Rawls makes it clear that arguments in the second stage are part of justification. Unless a conception of justice is shown to be stable in the second stage, "it is not a satisfactory political conception of justice and it must be in some ways revised." (PL:141, footnote 7) Besides, in a reply to Habermas's query about whether stability can add anything to the justification of a conception of justice, Rawls also states that "there is, then, no public justification for political society without a reasonable overlapping consensus, and such a justification also connects with the ideas of stability for the right reasons as well as of legitimacy." (PL:388-89) It shows that stability plays an important role in justification. We have no reason to believe that Rawls's restricting stability to the second stage is simply for practical considerations.

Why should stability be left to the second stage then? This is a puzzling

---

[6] Freeman, "Congruence and the Good of Justice," p.279.

question. If we concur with Rawls that stability is essential to justification, it is not clear why it should be ruled out from the first stage. To resolve this puzzle, we need to have a better understanding of Rawls's intention of dividing his theory into two stages.

## 2 Special Psychologies and Rational Choice

To begin with, we need to know what exactly the first stage refers to. According to Rawls, the first stage "gives the principles of justice that specify the fair terms of cooperation among citizens and specify when a society's basic institutions are just." (PL:133; also see PL:64) Moreover, "these two stages correspond to the two parts of the argument from the original position for the two principles of justice in *Theory*. In the first part the parties select principles without taking into account the effects of the special psychologies." (PL:140, footnote 7) This shows that the first stage refers to the original position in which principles of justice are chosen by rational parties behind the veil of ignorance. From a contractarian point of view, a conception of justice is more justifiable than another if it is chosen by mutually disinterested rational parties.

Details of the first-stage need not concern us for the moment. The thrust is whether stability is one of the considerations affecting the parties' rational decision making. If yes, we could say that it does provide justification for Rawls's principles in the first stage. The answer looks evident. In section 29 of *A Theory of Justice* under the heading of "the Main Grounds for the Two Principles of Justice," Rawls makes it clear that stability, together with the strains of commitment and self-respect, constitute the main grounds for rational parties to

favour justice as fairness. These factors "help to show that the two principles are an adequate minimum conception of justice in a situation of great uncertainty." (TJ:175/153 rev.) With regard to the specific role of stability, Rawls tells us that "a strong point in favour of a conception of justice is that it generates its own support." (TJ:177/154 rev.) Later on, this claim becomes even more explicit when Rawls states that "other things being equal, the persons in the original position will adopt the more stable scheme of principles." (TJ:455/398 rev.) This shows that stability is indeed an essential criterion to guide parties to *compare* and *choose* among conceptions of justice rather than merely *confirm* a specific conception that would have been consented to in advance. Rawls's claim that stability has no place in the first stage is inconsistent with this account.

Rawls replies that this reading has misunderstood the true nature of the two-stage procedure in his theory. He explains that the question of stability is concerned with the possible effect of special psychologies on the conception of justice, an issue which will only be brought to light in the second stage. Owing to the importance of this issue, let me quote a paragraph at length in which the later Rawls gives a clear account of this issue:

> We split the argument from the original position into two parts. In the first part, in which the principles of justice are *provisionally* chosen, the parties assume that the persons they represent are not moved by the special psychologies (or attitudes), as we called them. That is, the parties ignore persons' inclinations to be envious or spiteful, or to have a will to dominate or a tendency to be submissive, or to be peculiarly averse to uncertainty and risk...The second part of the argument concerns the question of the stability of justice as fairness...
>
> Together with the discussion of the special psychologies, the

> second part must take up the question whether in view of the general facts that characterize a democracy's political culture, and in particular the fact of reasonable pluralism, the political conception can be the focus of an overlapping consensus. We will consider how the question of stability leads to the idea of an overlapping consensus on a political conception of justice.[7]

Several points on this account are noteworthy. First of all, the main reason for dividing justice as fairness into two parts stems from Rawls's concern for special psychologies, especially the problems of envy and attitudes toward risk. In the first part the principles are derived on the supposition that these inclinations do not exist. The parties are mutually disinterested and their reasoning is not affected by envy or risk aversion. Since stability is set to check whether principles can be effectively applied to the basic structure without arousing envies to such a degree that social system becomes unworkable in the second stage, it plays no role in the first stage. This is not because doing so will lead to a Hobbesian conception of justice, but because the *circumstance of stability*, by definition, does not exist in the first stage. It is nothing more than a normative division of labour to deal with different issues of political justification.

Furthermore, the question of the overridingness of the sense of justice is also raised in this stage. It leads to the idea of an overlapping consensus. But we must note that this issue is different from the concern for special psychologies. A conception of justice, which may reduce people's envious attitude towards others, does not necessarily result in an overlapping consensus. For they address

---

[7] Rawls, *Justice as Fairness: a Restatement* ed. Erin Kelly (Cambridge, Mass: Cambridge University Press, 2001), pp.180-81. An almost identical account of the two-stage structure, except the idea of an overlapping consensus, can also be found in TJ: 144/124 rev., 504/441 rev., 530-31/465 rev..

essentially different issues. The former is to ask whether a well-ordered society regulated by a political conception will generate envy and other attitudes unfavorable to cooperation. The latter is concerned with the motivational priority of justice in a society of reasonable pluralism, to wit, how the sense of justice can be accepted by citizens as the regulative desire for their other interests resulting from their comprehensive doctrines. Bearing this distinction in mind is crucial when we assess how stability makes an influence on justice as fairness.

Now we can return to the question of why stability is excluded from the first stage. This section will only focus on special psychologies. An apparent reason is that Rawls does not want envy and risk aversion to affect the parties' rational deliberation. Their sole motive is assumed to that of advancing their conceptions of the good by winning for themselves the highest index of primary social goods. They have no interest in comparing their position with others. They do not "seek to maximize or minimize the difference between their success, and those of others" (TJ:144-45/125 rev.) Furthermore, as the parties are devoid of any special inclination toward risk, it is therefore rational to adopt the maximin rule as the criterion of rational decision making. They will select the conception whose worst outcome is superior to the worst outcome of any other alternative.

The exclusion of special psychologies seems to be closely related to the use of rational choice theory. There are two kinds of reasons. The first kind is a moral one. Rawls explains that if envy and knowledge of special psychologies were known to the parties, their choice would be affected by accidental contingencies. These contingencies should be avoided because they are generally regarded as morally undesirable. "The principles adopted should be invariant with respect to

variations in these inclinations for the same reason that we want them to hold irrespective of individual preferences and social circumstances." (TJ:530/464-465 rev.) In other words, they are as morally arbitrary as the differences in natural talents and social background. Excluding them from the original position can ensure that the agreement is made under reasonable conditions.

Nevertheless, this analogy is problematic. It is understandable that in justice as fairness the parties are not allowed to know their place in society or their fortune in the distribution of natural assets, for otherwise the ideals of fairness and equality would be compromised. Envy as a psychological inclination does not have such negative impact on the parties though. It will not affect the persons' equal status in the original position. Besides, if envy is arbitrary, so would be the assumption of mutual disinterest. From a moral point of view, it is not clear why the latter assumption would be any less arbitrary. So it is doubtful whether the exclusion of stability from the first stage can be ascribed to moral considerations.

More importantly, Rawls admits that the problem of envy should not be set aside and have its possible implication for a conception of justice neglected. For "these inclinations do exist and in some way they must be reckoned with." (TJ:530/465 rev.) What Rawls does is to leave the problem to the second stage rather than eliminate the issue entirely from his theory. He even says that if the conception of justice adopted in the first stage is found to arouse envy to such an extent that it brings social cooperation to its knees, then "the adoption of the conception of justice must be *reconsidered*." (TJ:531/465 rev.) This indicates that Rawls's treatment of special psychologies is entirely different from his attitude toward the arbitrary distribution of natural endowments and social positions. Any

moral objection based on the analogy would not be well grounded.

Rawls's second explanation is that the absence of special psychologies can provide an ideal environment for rational choice theory. For it can largely simplify the contractual situation and ensure that every person is fully rational in the original position. It is thus necessary that "the parties are not swayed by individual differences in these propensities, thereby avoiding the complications in the bargaining process that would result." (TJ:530/465 rev.) This is undeniably right. If the parties are moved by different motives, it is almost impossible to reach any rational agreement on a conception of justice. But why should simplification *per se* be such a strong reason to exclude special psychologies from the first stage? After all, what we search for is *right* principles. According to Rawls, however, the right principles are exactly those that could be unanimously consented to. Excluding "irrational" motives from the first stage is to ensure that the parties make their choice as fully rationally as possible. This is crucial because from a contractarian point of view, the question of justification is settled only if the principles are the result of collective rational choice. Since envy and risk aversion are inimical to rational reasoning, they are therefore kept out of the first stage. I believe this is the major reason for the two-stage design.

Let me elaborate this point a bit further. The basic assumption of rational choice theory is that people are rational. Rationality refers to economic rationality which means that a decision is rational if and only if it is the most effective means to realize one's informed end. It is sometimes also called means-end rationality. The end could be a person's interest, aims, or plan of life. Since the parties behind the veil of ignorance do not know their conceptions of the good, their common

goal is to secure as much primary social goods as possible. Moreover, the parties are mutually disinterested in one another's ends. They are presumed to be self-interested maximizers under the constraint of the original position. As Rawls puts it, "the persons in the original position are rational. In choosing between principles each tries as best he can to advance his interests.' (TJ:142/123 rev.)[8] This assumption of human motive warrants that principles of justice will be the result of rational choice.[9] In other words, the parties are not moved by benevolence. This is not only because benevolence is too demanding or too strong as a proper condition for the original position, but also because it would be incompatible with the use of rational choice theory.

Similarly, rational persons are presumed to be free from envy. By definition, envy is the propensity to view with hostility the greater advantages of others even though doing this may require us to give up something ourselves. Moreover, when other people aware of our envy, they may take a hostile attitude toward us. As a result, "envy is collectively disadvantageous: the individual who envies another is prepared to do things that make them both worse off." (TJ:532/466 rev.) Acting out of envy is therefore irrational from the point of view of means-end rationality. If envy was allowed to exist in the first stage, it would complicate the situation of rational bargaining and put the possibility of unanimous consent at risk.

The argument above begs a question: even if the parties are rational, how can it be guaranteed that they will prefer Rawls's principles to other alternatives, in

---

[8] The second sentence of the citation is deleted from the revised edition.

[9] It is artificial because Rawls says that "the motivation of the persons in the original position must not be confused with the motivation of persons in everyday life who accept the principles that would be chosen and who have the corresponding sense of justice." (TJ:148/128 rev.)

particular, the principle of average utility? After all, Rawls admits that if the parties behind the veil of ignorance adopt the principle of insufficient reason which assigns the same likelihood to all possible positions, it is quite natural for the parties to choose the principle of average utility. (TJ:165-6/143 rev.) Although it is riskier to reject Rawls's principles in favour of the principle of average utility, the parties may have a greater chance to gain more benefits should they have no aversion to risk. To prevent this, Rawls argues that it is most rational for the parties to adopt the maximin rule to guide their choice in the original position. In order to achieve this end, no knowledge of risk is available to the parties. "The essential thing is not to allow the principles chosen to depend on special attitudes toward risk. For this reason the veil of ignorance also rules out the knowledge of these inclinations: the parties do not know whether or not they have a characteristic aversion to taking chances." (TJ:172/149 rev.)

But Rawls seems to be self-defeating here. Does the maximin rule itself not reflect a special conservative attitude toward risk? On what grounds should the parties be prohibited from the knowledge of probabilities? In response, Rawls admits that the maximin rule is unusual and its application is only rational given the *unique features* of the original position. (TJ:172/149 rev., emphasis added) We now know that the parties' decision considerably hinges on the plausibility of those unique features. They determine the use of the maximin rule, which in turn determines Rawls's principles to be collectively chosen. The maximin rule plays the most important role in justifying Rawls's principles of justice. So in the coming sections our discussion will revolve around the maximin rule. As the discussion proceeds, the role of stability in Rawls's justificatory framework will gradually become clear. In the next section I will first make an objection to

Rawls's own claim that justice as fairness is essentially a social contract theory. This objection will carve out some space for my more critical claim that the maximin rule actually presupposes a moral argument.

## 3 The Rational and the Reasonable

It should now be clear that the main reason to exclude special psychologies from the first stage stems from Rawls's belief that rational choice theory is indispensable to moral justification. "In a contract theory all arguments, strictly speaking, are to be made in terms of what it would be rational to agree to in the original position." (TJ:75/65 rev.) Furthermore, "the theory of justice is a part, perhaps the most significant part, of the theory of rational choice." (TJ:16/15 rev.) The question is how Rawls can ensure that his principles would be chosen by the autonomous parties. There seems to be a tension between respecting the parties' voluntary choice and deriving the expected principles. On the one hand, Rawls purports to present his principles as a result of consent. In that case, its outcome is dependent on the contractors' rational agreement, but not subject to Rawls's own preference. A certain degree of indeterminacy is then inevitable. For if the principles are strictly pre-determined by the constraint of the contractual situation, the idea of consent would be redundant. One of the attractions of contract theory is that it expresses a message of respect for the autonomy and plurality of individuals. As Rawls states, "the merit of the contract terminology is that it conveys the idea that principles of justice may be conceived as principles that would be chosen by rational persons, and that in this way conceptions of justice

may be explained and justified." (TJ:16/14-15 rev.)[10] On the other hand, Rawls needs to take every necessary measure to ensure that his preferred principles would be the most rational candidates for adoption from the parties' point of view. To accomplish this, Rawls strives to provide a liberal egalitarian interpretation of the original position which best expresses the reasonable conditions imposed on the choice of principles. Rawls's hope is that the conditions can decisively lead the parties to choose his principles of justice. In his words, "the acknowledgement is the only choice consistent with the full description of the original position." (TJ: 121/104 rev.)

How to resolve this tension becomes a big issue for Rawls. The clue lies in justifying the maximin rule. If Rawls can show that the original position indeed allows room for the deliberation of different alternatives while the maximin rule is the only rational rule for decision making that eventually leads to his principles, then Rawls may achieve both ends without compromising his ideal of contractarianism. This is by no means an easy task, given that the tension reflects two different models of justification.

The first model is the application of rational choice theory to moral justification. Let us call it the *Rational Model.* According to Rawls, "in a contract theory all arguments, strictly speaking, are to be made in terms of what it would be rational to choose in the original position." (TJ: 75/65 rev.) Rational choice defines what justice is. But if justice is presented as an agreement of

---

[10] Of course I understand that this is a hypothetical rather than an actual contract. What is important is that as a contract theory, the decision of the parties must be presented as an autonomous choice made from their first-person point of view.

self-interested rational choosers, it gives readers a strong impression that it is a version of justice as mutual advantage. For there is no motivating force to appeal to but one's self-interest in acquiring the greatest amount of primary social goods. The rule for choosing a conception of justice is that it best advances each person's interest.[11]

Rawls however clarifies that this is not an accurate description of justice as fairness. This impression merely results from looking at but one of the elements of the original position. A proper understanding of justice as fairness should also take other conditions into account, especially the veil of ignorance which embodies the ideals of freedom and equality. The original position is an intricate design putting substantive moral judgments, formal conditions and general knowledge of human society together so as to ensure that rational decision is made under fair conditions. As Rawls puts it, "one way to look at the idea of the original position, therefore, is to see it as an expository device which sums up the meaning of these conditions and helps us to extract their consequences." (TJ:21/19 rev.) Viewing from this perspective, the justification of justice as fairness is no longer purely grounded on a model of rational choice for mutual advantage. For the decision-making is subject to a wider moral constraint which reflects our conception of moral persons as free and equal. The justificatory force is ultimately derived from this conception of the person rather than the rational choice of self-interested parties.

Rawls is surely right that the veil of ignorance has in effect forced every rational person in the original position to reason impartially and take the good of

---

[11] An incisive critique of justice as mutual advantage can be found in Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), pp. 28-46.

others into account. It is not a Hobbesian model of justice based on persons' actual bargaining power. This response falls short of dispelling doubts though. Even under the impartial conditions, Rawls's contract theory still requires that the arguments for his principles are presented as parties' autonomous choice. From the parties' rational point of view, their choice can only be justified in terms of its contribution to the advancement of their interest. Their decision is not grounded on any moral consideration. In this sense, justice as fairness does not appear to have any big difference from justice as mutual advantage. For example, after the veil of ignorance is lifted and parties' identities are revealed, there is nothing morally wrong or logically inconsistent for a person to reject the principles made in the original position and opt for a re-negotiation that can better promote their interests. As Dworkin aptly points out, "the fact, therefore, that a particular choice is in my interest at a particular time, under conditions of great uncertainty, is not a good argument for the fairness of enforcing that choice against me later under conditions of much greater knowledge."[12] The rational choice model of justification cannot explain the binding force of political principles when people come back to their real life.[13] This is a great challenge that a contract theory must answer. The later Rawls clarifies the role of rational choice in his theory as follows:

> These constraints are modeled in the original position and thereby imposed on the parties: their deliberations are subject,

---

[12] Dworkin, "The Original Position," in *Reading Rawls*, ed. Norma Daniels (Stanford, California: Stanford University Press, 1975), p.20.

[13] Rawls cannot say that the binding force comes from people's hypothetical agreement because, as Dworkin famously points out, "a hypothetical contract is not simply a pale form of an actual contract; it is no contract at all." "The Original Position," in *Reading Rawls*, p.18. Barry makes a similar critique in *Justice as Impartiality*, p.59.

> and subject *absolutely*, to the reasonable conditions the modeling of which makes the original position fair. *The Reasonable, then, is prior to the Rational, and this gives the priority of right.* Thus, it was an error in *TJ* (and very misleading one) to describe a theory of justice as part of the theory of rational choice, as on pp.16 and 583...There is no thought of trying to derive the content of justice within a framework that uses an idea of the rational as the sole normative idea.[14]

Rawls is well aware of the difficulty of rational choice as the sole basis of justification. His way out is to place more emphasis on the reasonable conditions to constrain the rational deliberation. The conditions circumscribe what alternatives would be put on the table for choice. More importantly, they embody the moral point of view. "In a contract theory, these moral conditions take the form of a description of the initial contractual situation." (TJ: 160/138 rev.) In response to the question about why people should take any interest in a hypothetical contract, Rawls's answer is that "the conditions embodied in the description of the original position are ones that we do in fact accept. Or if we do not, then perhaps we can be persuaded to do so by philosophical reflection." (TJ: 21/19 rev.)[15] The main grounds for justification lie in the normative prescription of the original position rather than the act of rational agreement. This is another model of justification which places the focus on the side of reasonableness. Let us call it the *Reasonable Model*.

---

[14] Rawls, "Justice as Fairness: Political not Metaphysical," in *Collected Papers* ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.401, footnote 20, my emphasis. To be fair, the message of "the reasonable is prior to the rational" abounds in *TJ* (pp.12/11rev., 18-19/16-17 rev., 21/19 rev., 120-21/104-05 rev., 446/392 rev., 516/453 rev., 585/512 rev.).

[15] Rawls repeats the same answer once again at the end of *TJ*. See TJ: 587/514 rev..

But if the rational is *absolutely* subject to the reasonable, and "the reasonable" is justified by independent reasons in advance, how much room will the original position leave for the rational parties' deliberation? When justice as fairness is no longer part of the rational choice theory, to what extent can it still be called a contract theory? Rawls holds that as a matter of fact there is no room for bargaining and deliberation under the reasonable constraints indeed. For instance, he states that "the aim of the contract approach is to establish that taken together they impose significant bounds on acceptable principles of justice. The ideal outcome would be that these conditions *determine* a *unique* set of principles." (TJ: 18/16 rev., my emphasis) And in commenting on the nature of his argument, Rawls stresses that "I should like to show that their acknowledgment is the *only choice* consistent with the full description of the original position. The argument aims eventually to be strictly *deductive*." (TJ: 121/104 rev., my emphasis) In other words, Rawls's conception of justice, which is supposed to be the result of unanimous agreement of a plurality of rational parties, turns out to be deductively determined by the constraints imposed on the original position. There is actually no justificatory force deriving from the parties' choice. More accurately speaking, the parties have little autonomy because their decisions are strictly limited by external constraints. Though in principle they are free to make any decisions, Rawls's principles are the only candidate consistent with the description of the reasonable constraints.

It is then doubtful whether justice as fairness is still a contract theory. In my view, a contract theory should at least meet the following three criteria. First, it must involve at least more than one party. Second, the principles of justice must be regarded as an outcome of the unforced and informed consent of the parties.

Third, the act of consent itself must be able to substantially account for the justifiability of a conception of justice.

The contractarian nature of justice as fairness is in doubt when we consider these criteria. First of all, owing to the veil of ignorance, the parties are ignorant of their personal identities. Their differences in social circumstances, natural endowment, and plans of life are intentionally concealed so as to ensure a fair negotiating environment. This, in effect, renders every person identical, reasoning in the same way. As Rawls reveals, "we can view the choice in the original position from the standpoint of one person selected at random. If anyone after due reflection prefers a conception of justice to another, then they all do, and a unanimous agreement can be reached." (TJ: 139/120 rev.) Though the original position is presented in a contract form, the nature of argument is one-person reasoning. There is neither bargaining nor exchange of views between parties. The picture of a plurality of persons coming together to deliberate a conception of justice for mutual advantage is an illusion. As Barry vividly describes it, "faced with identical information and reasoning in an identical fashion, they arrive at identical conclusions. We might as well talk of computers having the same program and fed the same input reaching an agreement."[16] As a result, Rawls's contract is a monologue rather than a dialogue.[17] It therefore fails to meet the first

---

[16] Barry, *Justice as Impartiality*, p.58.

[17] It could be argued that the situation a rational contractor is confronted with is that of an indefinite number of similar rational beings whose interests are diametrically opposed to his own. His problem is how to secure the largest amount of primary social goods for himself in circumstances where there are many others with the opposed aim of getting all the goods for themselves. It is therefore incorrect to say that the situation is a monologue. This is a possible interpretation. However, since the most rational thing for each to do in the circumstances is the same, the importance of plurality of persons is reduced to a minimum degree. I am indebted to John Charvet for reminding me of this alternative interpretation.

criterion of a social contract theory.

Furthermore, if the original position does not have any room for bargaining with others, and the so-called agreement is reduced to a one-person deliberation selected at random, we have reason to question whether it can be deemed an unforced and informed consent among individuals with different conceptions of the good. This does not mean that no constraint should be imposed on a contractual situation. As a matter of fact, any contract theory will have its own description of the initial situation, and the description will inevitably constrain the choices of individuals. But if a description is so rigid that the result is predetermined and the parties have little autonomy to make their choice, it should hardly be viewed as a consent-based theory. In this regard, Rawls's theory does not meet the second criterion either because the principles are the result of reasonable constraints rather than the consent of rational parties.

Finally, given that justice as fairness fails to satisfy the first two criteria, the idea of consent can no longer be said to be essential to justification. For the priority of right implies that the "consent" itself has little justificatory force. What ultimately matters is the reasonable conditions that substantially determine the content of political principles. In case there are difficulties in passing the test of rational choice, Rawls will revise the conditions of the original position to make it through.

Based on the above arguments, we can reasonably question whether justice as fairness can be called a contract theory. Rawls's intention is to make a proper division of labour between the reasonable and the rational to justify his principles

of justice. Both reasonableness and rationality are essential to his project. Without the reasonable constraint, rational choice cannot legitimize itself as a moral theory; without rational deliberation, the idea of contract becomes irrelevant. Although Rawls makes a great effort to strike a balance between them throughout his works, the tension remains. Plenty of textual evidence for both models can be found in *TJ*. This explains why critics often have different views about the ultimate grounds for the justification of justice as fairness. For example, we can easily understand Dworkin's challenge that the hypothetical contract does not play any substantive role in justification, and Barry's criticism that Rawls's version of rational choice theory confuses justice as mutual advantage and justice as impartiality. For Dworkin and Barry, the element of rational choice behind the veil of ignorance can in principle be put aside without damaging the integrity of Rawls's core moral principles. What matters fundamentally are those moral reasons that define the reasonable conditions of the original position. Rawls's acknowledgement of the priority of the reasonable over the rational actually concedes that the rational choice theory only plays a subordinate role in the original position.

Nevertheless, having taken all these arguments into account, Rawls can make a final rebuttal. He could argue that the reasonable conditions of the original position are only necessary, but not sufficient, premises to lead to his principles of justice. It is the maximin criterion that determines the parties' preference for Rawls's difference principle. If the rational parties did not adopt the conservative rule of decision making under the circumstance, Rawls's whole construction of the original position would not work. So in order to ensure that the difference principle will be the *only* choice made by the rational parties, the maximin rule must be justified as the *unique* and most *rational* rule under that circumstance. If

this argument succeeds, Rawls can then say that although his theory may not fully meet the criteria of contract theory, the idea of rational choice theory is still indispensable to his theory. After all, it can demonstrate that even from a single person's point of view, it is still up to the self-interested agent to decide which decision is more rational. Put it another way, the adoption of the maximin rule itself is based on rational choice. It is at this point that some room is left to the rational argument. Whether this strategy works, of course, depends upon the validity of Rawls's substantial argument for the maximin rule.

Here comes the core question: is it really necessary for the parties to adopt the extremely conservative maximin rule in the original position? Many critics disagree. For example, Harsanyi argues that it is more reasonable for the parties to adopt the principle of insufficient reason, which assigns the same probability to each possible place in a situation of uncertainty.[18] If so, it is highly likely that the parties would choose the principle of average utility rather than the difference principle. (TJ:165-66/143 rev.) Rawls responds that although maximin is not a suitable general guide for choices under uncertainty, it is reasonable for it to be applied in the original position in which knowledge of likelihood is impossible owing to the veil of ignorance. Nevertheless, the absence of empirical knowledge of probability does not mean that using subjective probabilities is unreasonable. On the contrary, using the maximin rule itself "is equivalent to assigning unity or near-unity probability to the possibility that one may end up as the worst-off

---

[18] Harsanyi calls this the 'equiprobability assumption'. "Can the Maximin Principle Serve as a Basis for Morality," in *John Rawls: Critical Assessment* vol.1, ed. Chandran Kukathas (London & New York: Routledge, 2003), p.223.

individual in society."[19] Furthermore, there is no obvious reason why the parties should be prohibited from having a general knowledge of their society. After all, this kind of general knowledge will not affect the impartiality of the original position. Rawls might say that risk aversion is another crucial factor in favour of the maximin rule. Yet this is the result of excluding the knowledge of special psychologies from the first stage. By the same token critics can question whether risk aversion itself is a consistent and rational attitude if one is ignorant of one's special attitude toward risk.

The pros and cons of the maximin rule have been extensively discussed. There seems no decisive reason to say that the maximin rule is the most rational strategy under uncertainty. The Bayesian principle of insufficient reason appears to be equally plausible. Rationality itself is not strong enough to support the maximin rule. If so, Rawls's whole project is at risk. Is there any other way to justify the maximin rule? I believe there is one. I will simply call it the *moral argument for maximin.* The main idea is that the fact of giving exclusive concern to the worst possible outcomes under alternative conceptions of justice expresses a particular moral point of view. It is not simply an issue of rational choice under uncertainty, or of psychological attitude towards risk. Rather, the parties accept the maximin rule because they take that moral point of view seriously. Contrary to Rawls's official account, they are not moved by self-interest. They have a sense of justice and strive to put forward moral arguments that other free and equal beings can reasonably accept.

---

[19] Harsanyi, "Can the Maximin Rule Serve as a Basis for Morality?" p.225.

At first sight the moral argument seems unbelievable. For it contradicts the central ideas of contractarianism. According to the standard account, the parties accept the maximin rule because it is the safest rule to help them protect their interest under the special situation of uncertainty. They have no incentive to take moral considerations into account. Moral reasons are only modelled into the reasonable conditions of the original position, but not directly applied by rational contractors. I am well aware that my argument, if sound, would have radical implication for justice as fairness. It will fundamentally change the nature of Rawls's theory. But there are some advantages of this argument.

First, since the parties are led by moral reasoning, the tension between the reasonable and the rational will be resolved. The constraint of the original position and the decision-making of parties form a coherent whole to express the ideal of fair cooperation between free and equal persons. It can avoid the charge that justice as fairness is a variant of the Hobbesian conception of justice as mutual advantage because the parties directly appeal to moral reasons to justify their choice.

Secondly, this moral argument can fill another gap between people's motivation in the original position and in a well-ordered society. Recall that in the original position the parties are presumed to be rational egoists while in the well-ordered society they are moved by the sense of justice. To avoid this motivational gap, Rawls reminds us that "the motivation of the persons in the original position must not be confused with the motivation of persons in everyday life who accept the principles that would be chosen and who have the corresponding sense of justice." (TJ:148/128 rev.) But this split of motivation begs

a big question: why must it be the case that after the veil of ignorance is lifted, rational persons will act in accordance with the sense of justice rather than self-interest? This brings us back to the basic problem of stability internal to the design of the original position. Although Rawls takes pains to bridge the gap by constructing psychological laws to account for citizens' moral development in a well-ordered society, he hardly resolves the problem. If the principles are presented as the result of consent among mutually disinterested persons, it is *unreasonable* to demand that the motive of self-interest be replaced by the sense of justice immediately after the veil is removed. It is not about whether they could psychologically make such a motivational shift, but whether it is logically consistent and morally reasonable to require it. To settle this, either one or the other motive has to be adjusted. Since Rawls takes moral stability seriously, a reasonable move is to allow the parties to be motivated by their sense of justice. The moral argument can exactly meet this requirement.

It might be argued that the moral argument would make the rational choice theory meaningless. Admittedly, introducing moral reasoning into the original position amounts to giving up the attempt of grounding justification on the rational agreement of self-interested individuals. But in my view, apart from the said advantages, this interpretation brings no real harm to Rawls. After all, in a strict sense justice as fairness cannot be described as a social contract theory. Moreover, the standard argument for the maximin rule is too weak to support Rawls's principles. We must search for other ways to justify the reasonableness of the maximin rule if we are to derive the principles from the framework of the

original position.[20] Of course to what extent this moral argument can make sense depends on the reasons offered. In the following two sections, I shall demonstrate that this interpretation has already been implied in Rawls's theory, or so I argue.

## 4 The First Moral Argument for the Maximin Rule

In this section I will start to illustrate the moral point of view behind the maximin rule. My general strategy is to reinterpret Rawls's arguments and show that the parties must appeal to some moral reasons should these arguments make any sense to them. Moral reasons are those that can be explained and justified with reference to a moral framework.[21] This section will first take up the issue of lexical priority of basic liberties.

To begin with, I would like to make a few remarks about Rawls's account of the maximin rule. Rawls holds that there are three special features of the original position that lead us to favour the maximin rule. First, it is impossible for the parties to have any knowledge of probability. The principle of insufficient reason is therefore inapplicable to the original position. Second, the parties would not take the risk of going for a further advantage when a satisfactory minimum is secured. Third, the rejected alternatives have consequences that the parties could not accept. In short, "the paradigm situation for following the maximin rule is when all three features are realized to the highest degree."(TJ:155/134 rev.)

---

[20] It is also possible that we give up the idea of the original position and find another device to justify Rawls's principles. The reasonableness of principles is then independent of the description of the original position.

[21] The precise meaning of moral reason need not bother us too much here. It suffices if we can demonstrate that the reasons for the maximin rule are not solely derived from rational self-interests. The reasons can refer to a conception of the moral person, an ideal of cooperation, or a commitment to equality.

With respect to the first condition, I have already argued that Rawls's appeal to the impossibility of calculating probabilities and risk aversion is not strong enough to turn down the principle of average utility. But with a more careful reading, we will notice that there is another normative reason urging the parties to take a conservative decision strategy, that is, "this supposition is plausible in view of the fundamental importance of the original agreement and the desire to have one's decision appear responsible to one's descendants who will be affected by it." (TJ:169/146 rev.) The parties understand that they are deciding the principles of justice which will have a fundamental impact on their life prospects and their descendants. It is the significance of choice that urges them to play safe. By contrast, the principle of average utility resulting from the principle of insufficient reason may require them to sacrifice their fundamental interest for the maximization of average utility under certain circumstances. The first condition actually depends on the third condition that other non-maximin alternatives have some intolerable outcomes.[22]

The second condition is closely related to the lexical priority of equal basic liberties. For Rawls, one of the weaknesses of utilitarianism is that it does not take individual right seriously. From the parties' point of view, it is said that they are not willing to try for greater gains at the expense of equal liberties. "The minimum assured by the two principles in lexical order is not one that the parties wish to jeopardize for the sake of greater economic and social advantages." (TJ:156/135 rev.) Thus, they adopt the maximin rule because it can more safely

---

[22] Scheffler makes a similar observation about this point. "Rawls and Utilitarianism," in *The Cambridge Companion to Rawls* ed. Samuel Freeman (Cambridge: Cambridge University Press, 2003), p.432.

protect their fundamental interests than alternatives. And these interests should never be sacrificed for other advantages. The second condition in effect also presupposes the third one that we should not choose those principles of justice which may have unacceptable outcomes when the veil of ignorance is lifted.

Therefore, the first two conditions actually presuppose the third condition. It all depends upon a comparison between Rawls's principles and other conceptions of justice concerning their possible impact on people's lives. Comparison needs criteria. Rawls then proposes a set of criteria and argues that in view of them, the parties would prefer his principles to other alternatives because the latter have some undesirable consequences that they are unwilling to bear, or the former have some advantages that other theories cannot provide. The focus is particularly on the comparison between Rawls's principles and utilitarianism. The central argument can be stated as follows: to justify Rawls's principles is to justify the maximin rule; to justify the maximin rule is to justify the criteria of comparison that the parties take seriously. Now what we need to do is to examine these criteria and see whether they express some moral ideals. If they do, my claim that the parties' reasoning is guided by moral considerations will be validated.

I will first discuss the argument for the lexical priority of equal basic liberties, which is closely related to the second condition aforementioned. Rawls's first principle states that "each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all." (TJ:302/266 rev.)[23] Furthermore, this principle enjoys a lexical priority over

---

[23] This principle has later been revised as "each person has an equal right to a fully adequate scheme of equal basic liberties which is compatible with a similar scheme of liberties for all." (PL:

the second principle, the principle of efficiency as well as the principle of utility. The right to equal basic liberties can only be restricted for the sake of liberty. It prohibits exchanges of liberties for other economic and social advantages. It expresses Rawls's commitment to individual rights which "even the welfare of society as a whole cannot override." (TJ:3/3 rev.) But from the point of view of rational persons behind the veil of ignorance, on what grounds are they willing to give an absolute priority to liberties over other primary goods? Why is it not rational for them to surrender some of their basic liberties, perhaps temporarily, in exchange for more material goods? To answer this question, we should bear in mind that the parties are guided by what they think is best for their own interest, not by any antecedent moral ideals or principles of rights. The justificatory reasons must be limited to those self-regarding ones.

In the first edition of *A Theory of Justice*, Rawls's answer is that as the conditions of civilization improve, our interest in exercising liberty will naturally increase while the marginal significance of material goods diminishes. Beyond some point, "it becomes and then remains irrational from the standpoint of the original position to acknowledge a lesser liberty for the sake of greater material means and amenities of office." (TJ:542)[24] This is because under favourable circumstances, the freedom to pursue our spiritual and cultural interests becomes more and more regulative in our plans of life. People have a fundamental interest in determining their conceptions of the good. Therefore, "the desire for liberty is the chief regulative interest that the parties must suppose they all will have in

---

291) My discussion will heavily focus on Rawls's arguments presented in "the Basic Liberties and their Priority," originally written for the Tanner Lecture in 1981, and later revised and collected in *Political Liberalism* as Lecture VIII .

[24] In the revised edition of *TJ*, Rawls has dropped this account and re-written this part.

common in due course." (TJ:543)[25]

The core argument for the precedence of liberties relies on an empirical account of human psychology. It presumes that rational persons will have a natural desire for the realization of freedom when socio- economic development reaches a certain level. But this assumption is doubtful. First of all, people with different characters and desires may have different views about the trade-off between liberty and economic interest. As Hart rightly points out, there is no ground for saying that a minor surrender of political liberties for a large increase in material welfare at some stage in the development of society is necessarily irrational.[26] As a matter of fact, even in advanced capitalist societies, many people are willing to sacrifice some of their liberties for the sake of greater material enjoyment.[27] Rawls overlooks that our ranking of different primary social goods is affected by a number of social factors such as culture, religion, education, and people's conceptions of the good. His argument for the absolute priority of liberty is therefore unsatisfactory.

Rawls may reply that under the special condition of uncertainty of the original position, it is only rational for the parties to opt for the principles whose worst consequence would be less undesirable than those of other alternatives.[28]

---

[25] This sentence has been deleted from the revised edition.

[26] H.L.A. Hart, "Rawls on Liberty and its Priority," in *Reading Rawls*, p.251. It should be noted that liberty is a concept of matter of degree. In many cases it is not an either/or situation between liberty and material goods.

[27] The political development of Hong Kong in the past decade, the city where I am living, is a case in point. Although many Hong Kong citizens favour democracy, they are, however, willing to make some compromise between political liberties and economic development when Beijing tells them that there must be a trade-off between these two goals.

[28] This point is raised by Hart, "Rawls on Liberty and its Priority," p.251.

The priority rule of liberty results from maximin reasoning. But the premise of this claim must be that the parties commonly treasure basic liberties as the *chief regulative interest* in their lives, to wit, the exercise of liberties occupies a higher-order place in their system of desires. The priority of a value implies its supreme significance in one's value system. Thus, to establish the lexical priority of basic liberties, we had better present the argument as a normative ideal rather than a psychological assertion. This is exactly the move Rawls makes in his revised edition of *A Theory of Justice* and his later works. He abandons the original argument and grounds the priority rule on an ideal conception of moral person.

Rawls re-formulates his arguments by saying that members of a well-ordered society have highest-order interests in realising their two moral powers. These powers are the capacity for a sense of justice and for a conception of the good. The former is the capacity to understand, to apply, and to be motivated by an effective desire to act from the principles of justice. The latter is the capacity to form, to revise, and to pursue a conception of what we regard as a good life. (PL:302) It is an interest of the highest order in a sense that it is supremely regulative as well as effective. This implies that "whenever circumstances are relevant to their fulfillment, these interests govern deliberation and conduct."[29] Rawls then prescribes that the rational parties, as representatives of moral persons, are likewise moved by these interests to secure the development and exercise of the moral powers. That being said, the argument for the priority of liberty

---

[29] Rawls, "Kantian Constructivism in Moral Theory," in *Collected Papers*, ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.312.

becomes rather straightforward because the parties are "simply trying to guarantee and to advance the requisite conditions for exercising the powers that characterize them as moral persons."[30] In other words, what Rawls needs to prove is that the basic liberties are necessary conditions to realize people's highest-order interests. Rawls thus remarks that:

> The persons in the original position are moved by a certain hierarchy of interests. They must first secure their highest-order interest and fundamental aims (only the general form of which is known to them), and this fact is reflected in the precedence they give to liberty; the acquisition of means that enable them to advance their other desires and ends has a subordinate place. (TJ:476, rev.)

The argument for the lexical priority of liberties is then completed. It is almost grounded on a deductive argument: the highest-order interest entails the priority of liberties. For example, freedom of conscience and freedom of thought are regarded as necessary conditions to develop people's capacity for a conception of the good.

It is apparent that this argument has fundamentally changed the nature of Rawls's theory. First of all, the parties' motive has been altered. These people are no longer moved by first-order interests. Instead, they are stipulated to have a common higher-order goal to search for a conception of justice that can realize their moral identity most effectively. Their reasoning is guided by a conception of moral personality. Rawls expressly acknowledges this paradigmatic shift by saying that "these revisions bring out that the basic liberties and their priority rest

---

[30] Rawls, "Kantian Constructivism in Moral Theory," p.315.

on a conception of the person that would be recognized as *liberal*, and *not*, as Hart thought, on considerations of *rational interest*." (PL:290, my emphasis) If so, it proves my moral argument for the maximin rule. The reason for the parties being not willing to sacrifice their liberties for the sake of material goods is that they take the exercise of moral powers as the regulative desire in their motivational set. They have a desire to be a liberal autonomous person.

How about those who do not accept this ideal? Rawls's reply is that "we expect and indeed want people to care about their liberties and opportunities in order to realize these powers, and we think they show *a lack of self-respect and weakness of character* in not doing so."[31] It indicates that in Rawls's mind the account of highest-order interests is a prescriptive moral ideal rather than an empirical description of the actual desire. The deliberation of parties can no longer claim to be morally neutral. The account of primary goods is also re-interpreted as the necessary means to enable human beings to realize their moral powers rather than simply to advance their different plans of life. As Rawls puts it in the "Preface for the Revised Edition", "primary goods are now characterized as what persons need in their status as free and equal citizens, and as normal and fully cooperating members of society over a complete life." (TJ:xiii, rev.)

Needless to say, how to justify this liberal conception of the person and the accompanying highest-order interest is a big issue. But this is not my concern here. My aim is to show that the parties would adopt a moral argument if they were required to use the maximin rule to justify the priority of equal basic liberties. The

---

[31] Rawls, "Kantian Constructivism in Moral Theory," p.315, my emphasis.

moral point of view is embodied in a liberal conception of the moral person. Interestingly, Rawls is reluctant to go that far. For admitting this would make the idea of rational choice meaningless. He thus says that the parties in the original position are only *rationally*, but not *fully autonomous*. The difference is that rational autonomy is acting solely from our capacity to be rational and from our conception of the good, while full autonomy includes not only the capacity to be rational but also the capacity to act in accordance with the fair terms of cooperation. Rawls stresses that although the parties are said to be moved by the highest-order interest in developing their moral powers, they are still regarded as rational self-interested beings. The priority of liberty is determined by its contribution to the well-being of persons that the parties represent. So, the revisionary introduction of highest-order interests will not affect the division of labour between the reasonable and the rational, and the principles can still be seen as a result of rational agreement. Rawls thus says, "the agreement in the original position on the two principles of justice must be an agreement founded on rationally autonomous reasons in this sense." (PL:307)

However, this defence is unsound. For the justification of highest-order interest in exercising moral powers cannot easily be translated to the satisfaction of self-interest.[32] Imagine that Peter is a rational egoist. His sole concern is how to gain as much material goods as possible to advance his plan of life. Peter does not deny the importance of some liberties because they are essential to realize his conception of the good. But differing from Rawls, he does not mind sacrificing a

---

[32] It should be noted that I am not claiming that the exercise of moral power is irrelevant to a person's well-being.

minor degree of his political liberties for a larger sum of economic benefits for various reasons. For instance, he may not regard the participation in political activity as an essential good. How can Rawls convince Peter that he is irrational to hold such a view? Rawls has little to say provided that this is Peter's deliberative and well-informed decision. If Rawls wants to persuade the parties to view the development of moral powers as their regulative goal, he must show that it is *unreasonable* (not irrational) for the parties not to take their moral personality seriously.[33] The frame of reference of reasonableness rests on the desirability of Rawls's account of the moral conception. As Hart acutely points out, the priority of liberty harbours a latent liberal ideal.[34] But ideal is different from interest. Whether a moral ideal is well-grounded does not depend on how useful it is to advance one's interest. In Rawls's case, his conception of the person is closely connected to the notion of fair social cooperation for reciprocity. Again, this notion is another ideal which expresses Rawls's commitment to fairness and equality.

To conclude, if the parties are moved by the highest-order interest in realising their moral powers, the justifying reason for choice must be moral in nature. The acceptance of the maximin rule is grounded on a conception of the moral person. One may argue that the moral conception does not need to apply to the rational parties directly. What we need is to model it into the original position as a reasonable constraint. But then it would be absurd that the parties should accept the exercise of the two moral powers as their regulative desire. Moreover, Rawls

---

[33] Of course whether the parties can be convinced by the reasons provided is another matter.

[34] Hart, "Rawls on Liberty and its Priority," p.252.

makes it clear that "the parties regard themselves as having certain fundamental interests that they must protect if they can; and that, as free persons, they have a highest-order interest in maintaining their liberty to revise and alter ends." (TJ:160 rev.) My argument therefore stands.

## 5 The Second Moral Argument for the Maximin Rule

Now I turn to examine the strains of commitment and stability. I will make two claims. First, I will show that they are actually the main grounds for the principles of justice in the original position. They provide major support for the adoption of the maximin rule in the first stage of justification. Rawls's claim that the problem of stability will not arise until the second stage is then wrong. Second, I will argue that the reasons behind the strains of commitment and stability embody a moral point of view as well. The parties are actually moved by moral reasons to adopt the maximin rule when they consider these issues.

To start with, we need to know how they are related to the maximin rule. The strains of commitment mean that in making their decision the parties will choose those principles that they can rely on one another to abide by. "They cannot enter into agreements that may have consequences they cannot accept." (TJ:176/153 rev.) Stability concerns whether citizens can have a sufficient sense of justice to act in accordance with the principles of justice. So both issues are concerned with the problem of motivation. Rawls stipulates that the rational parties should ask themselves whether they have sufficient motives to honour the principles of justice even if they belong to the least advantaged group when the veil is lifted. They must take this burden of commitment seriously. In view of the importance of

these considerations, it is therefore rational for them to prefer the maximin rule to other alternatives. Rawls makes this connection particularly clear at the beginning of Section 29, "the Main Grounds for the Two Principles of Justice":

> The arguments I shall adduce fit under the heuristic schema suggested by the reasons for following the maximin rule. That is, they help to show that the two principles are an adequate minimum conception of justice in a situation of great uncertainty. Any further advantages that might be won by the principle of utility, or whatever, are highly problematical, whereas the hardship if things turn out badly are intolerable. (TJ:175/153 rev.)

As we have shown in the foregoing section, among the three special conditions of the maximin rule, the third condition, namely the rejected alternative conceptions of justice whose consequences the parties could not accept, is the most essential and decisive one. Now we can see that the arguments of strains of commitment and stability are further elaborations of this condition. They provide extra support for the maximin rule. They have a direct and significant bearing on Rawls's principles of justice. Towards the end of the discussion in Section 29, Rawls even acknowledges that the limited information as to natural endowments and social status, the generality of principle, and universality of application are not enough to justify his principles of justice. They are only necessary, but not sufficient conditions. For without the adoption of the maximin rule, the parties may choose the average principle of utility. Rawls thus concludes that "the restrictions on valid undertakings as well as the publicity and finality conditions are *an essential part of the argument for the two principles.* I have discussed the

role of these constraints in connection with the strains of commitment and problem of stability." (TJ:183, emphasis added)[35] Furthermore, as Scheffler observes, the main ideas of the strains of commitment and stability set out in Section 29 are not fully developed until Part III of *Theory of Justice*. The central aim of Part III is to vindicate—through the discussion of rationality, the laws of moral psychology and the congruence argument and so on—that Rawls's principles would provide a satisfactory minimum, whereas utilitarianism might have consequences that the parties would find it difficult to bear. Therefore, Scheffler proposes that "the reasons for relying on the maximin rule...are actually the subject of much of the rest of the book. In effect, the 'maximin argument' functions as a *master* argument within which many of the book's more specific arguments are subsumed."[36]

If stability is indispensable to the justification of Rawls's principles in the original position, it is misleading for Rawls to say that it only arises at the second stage, and the arguments for stability "are not intended as justifying reasons." Nor is he right to claim that the function of stability is simply "checking whether the conception already adopted is a feasible one and not so unstable that some other choice might be better." (TJ:504/441 rev.) On the contrary, stability is one of the determining factors in the parties' decision. It is one of the "main arguments for the two principles" and helps "to show the two principles are an adequate minimum conception of justice in a situation of great uncertainty." (TJ:175/153 rev.) In Section 76 "The Problem of Relative Stability", Rawls makes a

---

[35] This citation is deleted from the revised edition.

[36] Samuel Scheffler, "Rawls and Utilitarianism," pp.435-36, my emphasis.

comparison between his principles and utilitarianism with reference to stability, and claims that "a decision in the original position depends on a comparison: other things equal, the preferred conception of justice is the most stable one." (TJ:498/436 rev.)

It is thus clear that stability plays an essential role in justifying Rawls's principles in the first stage. If this account is right, does it mean that envy should also be considered in the first stage? It does not. As I have shown before, the problems of envy and psychological attitude towards risk are not the only source of stability. Even if envy is excluded from the first stage, there could still be other reasons for taking stability into account. Otherwise, we cannot explain why the parties would take it as an essential ground for the maximin rule. We need an account to explain the role of stability in the first stage. That is the question we now turn to explore. The problem can perhaps be put this way: from the parties' standpoint, why would they adopt the strains of commitment and stability as normative constraints? According to Rawls, it is related to the formal constraints of finality and publicity respectively. I will discuss them in turn.

The condition of finality states that the principles adopted in the original position is the final court of appeal in practical reasoning. It specifies the totality of relevant considerations and their appropriate weight, and its requirements are decisive. Once the decision is made, there is no second chance for re-negotiation. The parties are aware of this constraint, and try to avoid those principles that they can adhere to only with great difficulty. Moreover, Rawls stresses that "when we enter an agreement we must be able to honour it even should the worst possibilities prove to be the case." (TJ:176/153 rev.) Suppose we accept this

burden of commitment as a necessary condition for justification. Here comes the question: why should we view the maximin rule and the resulting difference principle as the most appropriate candidate to meet this constraint?

Let us consider the following scenario suggested by Rawls: the parties are conducting a pair-wise comparison between the difference principle and the principle of average utility on the distribution of social and economic advantages. Both conceptions accept the priority of the principle of equal liberties and the principle of fair equality of opportunity. The extreme case of sacrificing someone's basic liberties for the sake of a greater good enjoyed by others does not exist.[37] This qualification can sharpen the comparison and help us see the justificatory force of the strains of commitment in a clearer way. In this circumstance, Rawls argues that the parties would still favour the difference principle rather than the principle of average utility. For the latter is psychologically too demanding. Adherence to it may exceed the capacity of human nature. From the point of view of the least advantaged, the utility principle asks them to view "the greater advantages of others who have more as a sufficient reason for having still lower prospects of life than otherwise they could be allowed."[38] By contrast, the difference principle assures them that inequalities will work to their greatest advantage.

At first sight, it is not clear why the principle of average utility is psychologically unbearable. After all, under this scheme their basic liberties and

---

[37] Rawls, "Some Reasons for the Maximin Criterion" in *Collected Papers* (Cambridge, Mass: Harvard University Press, 1999), pp.228-29.

[38] Rawls, "Some Reasons for the Maximin Criterion," p.230.

rights have been firmly secured. Even if it turns out that they are the least advantaged when the veil is lifted, they still have a fair chance to improve their economic situation because of the principle of fair equality of opportunity.[39] Moreover, following the principle of diminishing marginal utility, the parties have reason to believe that the principle of average utility will result in a rather egalitarian society. And in case they are not in the class of the worse-off, their economic prospect may be better than under the scheme of the difference principle. So it is not extremely risky to adopt the principle of insufficient reason and assign the equal probabilities to each possible position. The utility principle may satisfy the strains of commitment so that the argument for maximin based on the strains of commitment is indecisive.

There is, however, another interpretation of the strains of commitment.[40] It could be argued that although the principle of average utility may not be *psychologically intolerable*, they are *morally unbearable.* It is too demanding not because the absolute level of well-being arising from the application of the utilitarian principle is too low to command our respect; nor is it because we may have a higher chance to fall into the least advantaged group due to bad luck. Strictly speaking, any political principle may require substantial sacrifices of some people for others. From a libertarian point of view, for instance, the difference principle is extremely demanding because it requires people to share one another's fate, and treat the distribution of natural talents as a common asset. The better-off

---

[39] This consideration can be further strengthened by adding that the parties are all presumed to be normal and effective participants of cooperation.

[40] The following discussion has greatly benefited from Barry's incisive analysis in *Justice as Impartiality*, pp.61-67. But I disagree with Barry on the place of strains of commitment in the original position. I believe that it is a major argument for Rawls's principle in the first stage rather than in the second stage as Barry suggests.

may feel that they are forced labour for other's welfare in Rawls's cooperative scheme.[41] Furthermore, thanks to the priority of the difference principles over the principle of efficiency, the general level of well being of a Rawlsian society may even be lower than a utilitarian one. It is thus misleading to explain the strains of commitment in terms of psychological propensity.

Rawls's real argument is actually that it is *unfair* for utilitarianism to require someone to sacrifice their life prospects for the greater advantages of others if we regard each other as equal participants of cooperation for reciprocity. As he claims, "when society is conceived as a system of cooperation designed to advance the good of its members, it seems quite incredible that some citizens should be expected...to accept lower prospects of life for the sake of others." (TJ:178/155 rev.) They find it unacceptable because they have already accepted a specific conception of society as fair cooperation between free and equal citizens who have distinct conceptions of the good. If the parties are all strong communitarians and view cooperation as a shared project for the common good, they may not find utilitarianism as intolerable as Rawls assumes. In other words, utilitarianism is too demanding mainly because it fails to take the separateness of individuals seriously, and places unfair burdens on some people in cooperation. It is unreasonable from a moral point of view irrespective of whether it is psychologically demanding or not. It is a moral, but not psychological argument as it can be made sense of only if the parties have already endorsed a moral commitment to equality and fairness in advance.[42] The implicit reasons determining what principles can meet the test

---

[41] For example, See *Nozick, Anarchy, State and Utopia*, (New York: Basic Books, 1974), Chap. 7.

[42] Joshua Cohen has an excellent analysis of the egalitarian implication of the maximin rule in ""Democratic Equality," *Ethics* 99, pp.727-51.

of the strains of commitment actually express a particular liberal point of view, which in turn determines why the parties would adopt the maximin rule.

Indeed, Rawls's article "Some Reasons for the Maximin Rule" published in 1974 confirms my argument. In response to his critics, Rawls admits that his previous arguments for the maximin rule, including considerable risk-aversion, less demanding information requirements, greater suitability as a public principles and weaker strains of commitment, are not decisive by themselves. There must be other more *compelling* reasons for the maximin rule. He suggests that "the aspirations of free and equal personality point directly to the maximin criterion."[43] His idea is that since citizens view themselves as free and equal persons, they do not endorse any claim that one deserves one's place in the distribution of natural talents. Furthermore, the distribution of talents is viewed in some respects as a collective asset. It follows that the maximin rule is the most appropriate candidate to enshrine this moral ideal. A principle of justice can claim to be justified if and only if it could be reasonably accepted by free and equal moral persons, including those least advantaged. For every one has an equal power regardless of their differences in social and natural advantages. Rawls thus concludes:

> Provided the maximin criterion is satisfied, these relations may be preserved: inequalities are to everyone's advantage and those able to gain from their good fortune do so in ways agreeable to those less favored. *Meeting this burden of proof reflects the value of equality.*[44]

What Rawls calls the compelling reason for the maximin rule turns out to be

---

[43] Rawls, "Some Reasons for the Maximin Rule," p.230.

[44] Rawls, "Some Reasons for the Maximin Rule," p.231, my emphasis.

the value of equality. This is obviously a moral argument. Since the maximin rule is adopted by the parties themselves to guide their reasoning, they must accept the moral argument beforehand.[45] In this case, the distinction between the reasonable and the rational in the original position is no longer important. The parties are moved by a consideration of equality. If so, Rawls's principles of justice are deductively derived from his moral premise of equality, and the idea of a contract between rational self-interested persons becomes redundant.

Now let us turn to the issue of stability. What interests us is why stability would be considered the main ground for the principles of justice. This question has been discussed thoroughly in the previous two chapters. We have already known that Rawls's conception of stability refers to moral stability which is essentially concerned with the motivational priority of the sense of justice. But from the point of view of self-interested parties, why are they interested in this question? A natural answer seems to be that social stability can improve everyone's prospect if all comply with the principles of justice in the well-ordered society. Stability matters because it is a social virtue for effective cooperation. But this seemingly reasonable answer is not what Rawls holds.

First of all, Rawls points out that it is not irrational for a person to be a first-person and free-rider egoist when they come back to society from the original position. "In everyday life an individual, if he is so inclined, can sometimes win even greater benefits for himself by taking advantage of the cooperative efforts of

---

[45] One may argue that this argument only applies to citizens in the well-ordered society, but not to rational persons in the original position. This does not make sense because the maximin rule is only applicable to a special situation like the original position.

others." (TJ:497/435 rev.) Acting justly may not be each person's best interest in the real world. The parties should not be surprised by this possible consequence because they themselves are solely motivated by self-interest. The sense of justice has no effect on their deliberation. So, as Barry suggests, a more rational policy for the parties is to agree on an effective mechanism of enforcement to "prevent backsliding and to provide those who accept the principles with assurance that others will play their part by, for example, paying their taxes."[46] Stability, in this sense, is a purely practical matter. Its sole task is to find means of persuasion or enforcement to ensure compliance with the principles that have been independently worked out as reasonable. This is not Rawls's account of stability though. Rawls expressly indicates that his conception of stability is closely connected to the desirability of a conception of justice. As he puts it, "finding a stable conception is not simply a matter of avoiding futility. Rather, what counts is the kind of stability, the nature of the forces that secure it." (PL:142) The force of stability must originate from a reasonable conception of justice.

Here comes the critical question: from the standpoint of rational self-interested individuals, why should they impose the normative constraint on themselves that a principle of justice will not be justified unless it warrants the motivational priority of the sense of justice? It does not matter for the moment whether such a principle can be found. The real challenge lies in explaining why these self-interested persons are willing to take the priority of moral motivation so seriously. There is a motive inconsistency here. The only possible reason is that

---

[46] Barry, *Justice as Impartiality*, p.62.

the motivational priority is instrumental to social stability.[47] This account has been rejected by Rawls himself. However, he could not say that we should not confuse the motive of the contracting parties with the motive of citizens in a well-ordered society. For what the parties ask is exactly why they should draw this distinction and give priority to the sense of justice in case justice is in conflict with their self-interest.

In my view, there is no way to break the deadlock if the parties remain egoistic. An egoist cannot be persuaded to be a moral person by non-instrumental reason. If he wants to be a just person, he must commit to it. So the alternative is to change the motive of the parties from mutual disinterest to a desire to act justly. Once this change is made, it would not be unusual for the parties to take moral stability as an important consideration for the maximin rule. They care about moral stability because they have adopted a moral standpoint that a justifiable conception of justice must be able to offer reasons for the motivational priority of the sense of justice. As Rawls claims, "justice as fairness is not reasonable in the first place unless in a suitable way it can win its support by addressing each citizen's reason."(PL:143) My argument is also consistent with Rawls's claim that the parties are moved by the highest-order interest in developing their two moral powers. Both lead to the same conclusion that the parties should have a moral sentiment to reason and to act morally.

Some conclusions of the previous two sections are in order. First, I have shown that stability is indeed a justifying reason for Rawls's principles in the first

---

[47] I have refuted this account in Chapter 1.

stage. It offers substantive support for the maximin rule and forms a major ground for the principles of justice. Rawls is mistaken to say that stability is merely an issue concerning the feasibility of principles carried out in the second stage. Second, after examining the arguments for the priority of liberties, the strains of commitment and stability, we can see that Rawls's principles of justice are in fact grounded on moral reasons. To vindicate the reasonableness of the maximin rule, the parties must appeal to a liberal conception of moral persons as free and equal, and as having a higher-order interest in developing their moral power. This argument stands to overrule the whole idea of grounding justification on rational choice theory. But I do not think this interpretation will weaken the desirability of Rawls's principles. On the contrary, it can make the argument more consistent and powerful. What ultimately matters are those moral ideals behind the principles. It is unwise to invoke the rational choice strategy to explain the moral attractiveness of these principles. For this strategy is starkly inadequate in resolving the motivational gap, and justifying the regulative priority of the sense of justice over other desires. Rawls is well aware of this problem inherent in his rational choice argument. This partly explains why he needs the stability argument to bridge this gap. The issue of moral motivation is of utmost importance in Rawls's theory.

That being said, my argument leaves a question unsolved. If the stability question has already been answered in the first stage, why does justice as fairness need the second stage? In particular, if the parties are said to have the highest order interest in developing their capacity for a sense of justice, and to take the strains of commitment and stability into consideration in the first stage, what is the point of Rawls adding that "the argument for the principles of justice is not complete until the principles selected in the first part are shown in the second part

to be sufficiently stable." (PL:141, footnote 7) This question puzzles many of Rawls's critics. And a reasonable answer is necessary because the problems of congruence of the right and the good, and of an overlapping consensus are being dealt with in the second stage. In the following sections, we will continue to explore the place of stability in the second stage of justice as fairness.

## 6 The Need for the Second Stage

Our discussion will focus on three questions. First, what is the major concern of the second stage? Second, where does the second stage take place? Third, what important implications can we draw from the answers to the first two questions?

To begin with, I would like to make two preliminary remarks. First, stability in Rawls's context is always an issue concerning the motivational priority of the sense of justice. So when Rawls stipulates that we need a separate stage to handle the question of stability, it implies that there must be some important issues concerning the priority of the sense of justice that the first stage is unable to answer fully. What those issues are will be the key to appreciate the role of the second stage. Moreover, we should note that the principles of justice have already been established in the first stage. This does not surprise us because this is the task of the parties in the original position behind the veil of ignorance. Therefore, the main concern of the second stage is not whether Rawls's principles would be chosen by the parties. That question has been settled, and the considerations of the second stage "do not determine the initial acknowledgement of principles in the first part of the argument, but confirm it." (TJ:580/508 rev.) This feature is even further stressed in Rawls' s later philosophy. He says:

> In the first stage it [justice as fairness] is worked out as a freestanding political (but of course moral) conception for the basic structure of society. Only with this done and its content—its principles of justice and ideals—provisionally on hand do we take up, in the second stage, the problem whether justice as fairness is sufficiently stable. " (PL:140-141)

Then why should the justification of justice as fairness need the second stage? A straightforward answer is that the problem of motivational priority of justice will be resolved only in this stage. As I have thoroughly argued in the preceding chapter, a justifiable conception of justice must demonstrate its ability to motivate rational people to give priority to justice over other desires. This is related to Rawls's understanding of practical reasoning and his internalist position. Rawls's hope is to show that the sense of justice specified by his principles of justice can be a regulative desire in people's rational plans of life even judging from their personal perspective. Only if this is done, will citizens living in the well-ordered society have sufficient reasons to abide by the principles of justice. Rawls aims to argue for the good of the sense of justice in the second stage. As he puts it,

> We want to know whether having and maintaining a sense of justice is a good (in the thin sense) for persons who are members of a well-ordered society... And if within the thin theory it turns out that having a sense of justice is indeed a good, then a well-ordered society is as stable as one can hope for. Not only does it generate its own supportive moral attitudes, but *these attitudes are desirable from the standpoint of rational persons who have them when they assess their situation independently from the constraints of justice. This match between justice and goodness I refer to as congruence.* (TJ:398-99/350 rev., my emphasis)

The second stage is set to deal with the problem of congruence. For Rawls, "whether these two points of view are congruent is likely to be a crucial factor in determining stability." (TJ: 567/497 rev.) The idea of congruence indicates that there are two different points of view in guiding our actions. One is the standpoint of justice; another is the standpoint of goodness. The former is characterized by the arguments of the first stage. It embodies the idea of fair cooperation between free and equal moral persons. It requires us to look at each other from an impartial perspective abstracted from our personal identities and conceptions of the good. All participants are morally equal, and therefore have an equal right to decide the principles of justice. They do not look at the social order from their situation but take up a common point of view that everyone can adopt on an equal footing, which in turn defines what justice is. This is what Rawls calls the first stage argument.

Nevertheless, this standpoint alone cannot fully explain the practical reason for our action in real life. When people leave the original position and move to the well-ordered society, they will know their distinct conceptions of the good. Without the constraint of the veil of ignorance, many of them may be moved by their particular attachments and interests. The standpoint of goodness comes into play at this stage. We recognize that the sense of justice is just one of the many desires in an agent's subjective motivational set. It does not necessarily override other desires in practical reasoning. The priority of the sense of justice is not a foregone conclusion even in a well-ordered society. It must be substantively argued for. Rawls believes that the solution rests on the congruence of justice and goodness. The central idea is to render the desire to act justly regulative in a rational plan of life. There will then be no conflict between the two standpoints;

citizens will find the sense of justice desirable even "from the standpoint of rational persons who have them when they assess their situation independently from the constraints of justice." (TJ:399/350 rev.) When the congruence argument succeeds, justice as fairness will be as stable as one can hope for.

It is now clear that the function of the second stage is to ensure the priority of the sense of justice. Without this stage, the justification is incomplete because the impartial standpoint specified in the first stage is not enough to demonstrate why it is rational for citizens to give the sense of justice a definite priority judging from their viewpoint of goodness. As a reason/motive internalist, Rawls holds that the priority problem does not "rely on the doctrine of the pure conscientious act." (TJ: 569/499 rev.) It is against this background that we can understand why Rawls says that the considerations of the second stage are to *confirm* the initial choice of principles of justice in the first stage. Its success will show that "our nature is such as to allow the original choice to be carried through." (TJ: 580/508 rev.)

One might demand, legitimately, at this point why, if the question of stability has already been considered by the parties in the original position, it should be taken up again in the second stage? Rawls does not give any explicit answer to this question. But if my above argument is sound, it is fairly obvious that the argument for stability cannot be completed in the first stage. As argued in the previous section, the parties take the concern of stability as the main reason to adopt the maximin rule because they know that one defining feature of a well-ordered society is the willingness of citizens to act justly. Whether a conception of justice can generate its own support therefore becomes an important criterion of their decision. Since a general knowledge of moral psychology is

available, they can compare the relative stability of different conceptions of justice. Yet the decision made after the comparison is not decisive in assuring the priority of the sense of justice. For the viewpoint of goodness does not enter their deliberation. There is no way for them to judge whether the sense of justice characterized by a specific conception of justice can be a regulative good of their plans of life. They simply lack such information and perspective. Even if justice as fairness is shown to be more stable than the principle of utility, it does not warrant its motivational priority over other conceptions of justice. As Rawls remarks, "congruence is not a foregone conclusion even in a well-ordered society. We must verify it." (TJ:567/497 rev.) Therefore, both stages are essential to Rawls's project.

We can now see that the arguments for stability that appeared in Section 29 "the Main Grounds for the Two Principles of Justice" and Section 76 "the Problem of Relative Stability" mainly serve the first stage in determining the principles of justice, while the congruence argument of Chapter 9 "the Good of Justice" is for the second stage.[48] Unfortunately, Rawls himself does not make such a distinction; he misleads his readers into believing that stability is only a matter concerning the feasibility of principles of justice in the second stage. It ceases to be surprising why so much criticism and misunderstanding surround Rawls's account of stability.

There is one more point about the necessity of the second stage. Recall that a

---

[48] Scheffler makes a similar observation about the close connection of the stability argument of Sections 29 & 76 in justifying Rawls's principles. He points out that it is misleading for Rawls to say that stability is not a justifying reason for his principles. What Scheffler overlooks is the two-stage structure and the importance of the second stage in Rawls's theory. See Scheffler, "Rawls and Utilitarianism," p.455.

motivational gap exists between the original position and the well-ordered society. The parties are assumed to be motivated solely by self-interest. To avoid justice as fairness being an egoistic theory, Rawls must demonstrate that the parties can develop an effective moral motive to comply with the principles of justice when they return to their ordinary life. But this argument cannot be done in the first stage. Although the parties know that they will have a sense of justice in the well-ordered society, they do not know how effective it is after their plans of life have been revealed to them. As a result, the second stage argument is necessary to bridge the gap.

Having discussed the role of second stage, now we come to another crucial question: where does the second stage of stability take place? It seems obvious that this stage arises only after the veil of ignorance is lifted and the parties have full knowledge of their conceptions of the good. If persons are still situated in the original position behind the veil of ignorance, there is no basis for the congruence argument to proceed. Congruence must presuppose the existence of two standpoints. But in the original position, only the standpoint of justice is present. That is why Rawls says that in considering the match of two standpoints, people "assess their situation independently from the constraints of justice." (TJ:399/350 rev.) Furthermore, the ultimate concern of congruence is "whether the regulative desire to adopt the standpoint of justice belongs to a person's own good when viewed in the light of the thin theory *with no restrictions on information.*" (TJ: 567/497 rev., my emphasis) As for explaining the different nature of the two standpoints, Rawls reminds us again that "the requisite match exists between the principles of justice that would be agreed to in the absence of information and the principles of rational choice that are not chosen at all and *applied with full*

*knowledge.*" (TJ:514/451 rev., my emphasis) It is then clear that the argument for congruence takes place only after the thick veil is removed. People make their decision with full knowledge of their conceptions of the good. This observation is further confirmed by Rawls's argument for political liberalism. Though the later Rawls replaces the congruence argument with the idea of an overlapping consensus, the two-stage justificatory structure remains basically unchanged. Therefore, the problem of an overlapping consensus will only be taken up in the second stage and in such a consensus, "the reasonable doctrines endorse the political conception, each from its own point of view." (PL:134) It indicates that people must be allowed to know their conceptions of the good before they decide whether a consensus among reasonable doctrines is possible or not.

We now have a very different picture of Rawls's theory. First of all, if the congruence argument occurs outside the original position and the second stage is indispensable, then the rational choice made in the first stage is no longer final in Rawls's theory. The justifiability of a conception of justice is not finally determined by the choice of the original position. We must wait and see whether it will win the congruence argument, or become the focus of an overlapping consensus. Only when it is shown to do so, will moral stability be secured. In other words, a full justification of justice as fairness must go beyond the original position and rely on the success of the second stage argument for the priority of the sense of justice. Stability in this regard is undoubtedly a decisive criterion for the desirability of justice. This sheds light on the claim that "the argument for the principles of justice is not complete until the principles selected in the first part are shown in the second part to be sufficiently stable." (PL: 141, footnote7) But as the second stage goes beyond the original position, Rawls can no longer appeal to

an impartial and common standpoint to realize moral stability. He must provide separate procedure and substantive reasons to justify the overridingness of the sense of justice in the second stage. This is what I am going to examine in the next chapter.

## 7 Conclusion

This chapter has made several important claims which, if sound, would provide a fundamentally different picture of Rawls's theory of justice. First, I have argued that stability plays an essential role in both stages of justice as fairness. It helps the parties make their decision in the original position on the one hand, and accounts for congruence of the right and the good in the second stage on the other. Rawls's claim that stability is only considered in the second stage is thus misleading. Second, I have shown that the second stage is indispensable to the justifiability of Rawls's principles because the motivational priority can only be determined in the second stage. It is therefore incorrect for Rawls to say in *A Theory of Justice* that the second stage is only concerned with feasibility without any effect on principles derived from the first stage. Unless the principles are shown to be stable in the second stage, quoting the later Rawls's own words, "it must be in some way revised." (PL:141) Third, I have contended that the idea of contract based on rational choice has only a minor place in Rawls's theory. The grounds for the maximin rule, which embodies Rawls's ideal of the moral person and of liberal commitment to justification, are actually based on moral reasons. The assumption of mutual disinterest thus weakens Rawls's moral argument and leaves a motivational gap that makes moral stability almost unthinkable. This is why I have reservations about the division of labour between the reasonable and

the rational, two ideas expressing two incompatible models of justification. Finally, though my arguments raise strong objections to Rawls's propositions, it does not mean that the idea of stability is unimportant. Quite the opposite, what I do is to make the idea of stability more consistent and essential in the complex justificatory structure of justice as fairness. With the place of stability settled, we can proceed to examine the second stage arguments for stability in Rawls's early and later philosophy, to wit, the congruence and the overlapping consensus arguments in the coming chapters.

## CHAPTER 4

# CONGRUENCE, RATIONALITY AND TELEOLOGY

This chapter sets out to examine Rawls's argument for congruence. I shall first explicate the main ideas of congruence and then show that its main ground lies in a Kantian interpretation of justice as fairness. I argue that this interpretation has turned Rawls into a liberal perfectionist within a classical teleological framework. This position is, however, inconsistent with Rawls's desire-based conception of deliberative rationality. For this conception of rationality does not warrant that rational persons would necessarily accept a Kantian conception of the good. I shall conclude that it is this internal inconsistency which makes Rawls's argument for congruence fundamentally flawed and accounts for his philosophical turn to political liberalism.[1]

This is a long chapter, not only in pages. The whole essay consists of eight sections, which are arranged as follows. Section 1 will introduce the main idea of congruence and shows its connection to moral stability. Since the need of congruence has been challenged by Barry as based on a misunderstanding of Ross's doctrine of the purely conscientious act, Section 2 sets out to compare Ross's doctrine with Rawls's theory of practical reason for action. I argue that deeply influenced by Foot, Rawls believes that the good of the sense of justice is required to justify the motivational priority of justice. In Sections 3 and 4, I start

---

[1] As Rawls explains, "to understand the nature and extent of the differences, one must see them as arising from trying to resolve a serious problem internal to justice as fairness, namely from the fact that the account of stability in Part III of *Theory* is not consistent with the view as a whole. I believe all differences are consequences of removing that inconsistency. Otherwise these lectures take the structure and content of *Theory* to remain substantially the same." (PL:xvii-xviii)

to examine Rawls's free-rider argument and social union argument for congruence. I shall argue that both arguments fail. The free-rider argument fails because an egoist would not act on moral reasons. The social union argument fails because Rawls's account of social union is insufficient to establish a shared final end for the priority of the sense of justice. What Rawls ultimately relies on is the Kantian interpretation argument, which is the focus of Section 5. Through careful exposition, I shall argue that this interpretation is a type of perfectionism grounded in a conception of human nature as free and equal rational being. Moreover, my discussion about the second-order conception of personal autonomy in turn brings out an important distinction between neutral freedom and good freedom primarily suggested by Sidgwick in his commentary on Kant's moral philosophy. In Section 6, I suggest that the Kantian interpretation embodies both conceptions of freedom. However, these conceptions are internally incompatible with each other, and the pursuit of congruence requires Rawls to keep only the conception of good freedom. In Section 7, I shall show that the conception of good freedom has further revealed the Kantian interpretation as a classical teleological theory. The dichotomy between modern teleology and deontology obscures the teleological nature of Rawls's theory. Lastly, in Section 8, I maintain that there is an inconsistency between Rawls's teleology and his account of deliberative rationality, which results in the failure of congruence. My argument shall then shed light on Rawls's later development of political liberalism.

## 1 The Idea of Congruence

To begin with, I shall recapitulate the main ideas of congruence. Congruence

refers to the convergence of two distinct standpoints, namely the standpoint of justice and the standpoint of goodness. The former is defined by principles chosen in the original position of equality by free and rational individuals under the reasonable constraints, while the latter is defined by the successful execution of people's plans of life consistent with the criteria of deliberative rationality. As rational moral beings, we are guided by both standpoints in practical reasoning. The moral standpoint dictates us to do what justice requires while the rational standpoint moves us to realize our informed desire. They form the basis from which "institutions, actions and plans of life can be assessed." (TJ: 567/496-497 rev.) Both standpoints are legitimate, prescriptive and action-guiding.

Nevertheless, these standpoints do not always coincide with each other. An action which is right is not necessarily truly good for a rational agent. When the standpoints diverge from each other, the question of why it is rational for a person to act justly arises. According to Foot, if we cannot commend justice to people as a good, then "justice can no longer be recommended as a virtue."[2] As a result, the motivational priority of justice will be unfounded. The existence of two standpoints and the possibility of conflict between them set the background for congruence.

To argue for the overridingness of the sense of justice, Rawls suggests that under certain ideal condition, these two perspectives can converge. In that case, rational persons would have sufficient motive to abide by principles of justice. The basic question of congruence, according to Freeman, is this:

---

[2] Philippa Foot, "Moral Beliefs" in *Virtues and Vices* (Oxford: Oxford University Press, 2002), p.125.

> Is it rational in a well-ordered society of justice as fairness for persons to affirm individually, from the point of view deliberative rationality, the principles of justice they would rationally agree to when they take up the public perspective of justice?[3]

Rawls's answer is that in a well-ordered society regulated by justice as fairness, the sense of justice will be perceived as a regulative good by rational people. In such a society, Rawls believes that:

> Not only does it generate its own supportive moral attitudes, but these attitudes are desirable from the standpoint of rational persons who have them when they assess their situation *independently from the constraints of justice.* (TJ:398-399/350 rev., my emphasis)

Thus congruence is a decisive factor in determining stability. It represents a harmony between the moral life and the good life. The demand of justice is not regarded as an external constraint imposed on a rational free agent. Instead, it is presented as an important part of a conception of the good life. Congruence expresses a distinctive ethical view that practical reasons for action must stem from one's conception of the good. To establish the priority of the sense of justice, we need to show how it can occupy a central place in an agent's "subjective motivational set."[4] This view has been seriously challenged by Barry. Barry believes that the idea of congruence has in effect denied the claim that recognising something to be right is sufficient to motivate right action.[5] Put another way, an

---

[3] Freeman, "Congruence and the Good of Justice," in *The Cambridge Companion to Rawls* (Cambridge: Cambridge University Press, 2003), p.285

[4] This term is borrowed from Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), pp.101-13.

[5] Barry, "John Rawls and the Search for Stability," *Ethics* 105, (1995), p.884.

agent would not do what is right simply out of a sense of duty. Barry suggests that Rawls's view stems from his rejection of the doctrine of the pure conscientious act held by Ross, according to which "the highest moral motive is the desire to do what is right, and just simply because it is right and just, no other description being appropriate." (TJ:477/418 rev.) Having rejected this peculiar doctrine, according to Barry, "Rawls commits himself in Chapter 9 of *A Theory of Justice* to the ancient doctrine that no act can be regarded as rational unless it is for the good of the agent to perform it." [6]

Barry argues that Rawls makes an apparent mistake here because the desire to act morally for the sake of justice is a widespread disposition in society. People simply fulfil their duty out of their sense of duty and not for some independently defined good. Furthermore, the idea of congruence is unnecessary and absurd. It is unnecessary because Rawls need not rely on it to solve the problem of stability. In Chapter 8 of *A Theory of Justice*, Rawls has already shown that following the three laws of moral psychology people will develop an effective sense of justice to do what justice requires. Barry believes that this widely shared moral disposition is good enough to maintain a stable society without recourse to the demanding congruence argument. Rawls's claim is absurd because it will dissolve an important distinction between the right and the good in our practical reasoning, to wit, that we are capable of doing something right while believing that what morality forbids is good for us. Barry offers an example to illustrate this point:

> Suppose that I form the view that it would contribute to my good to take a trip around the world, and that I then find that

---

[6] Barry, "John Rawls and the Search for Stability," p.885.

> this would cost more than my resources permit...Instead of simply concluding that I cannot justly take the trip (while continuing to believe that taking it would be for my good), I am told by Rawls that I must somehow persuade myself that it would not be for my good at all. *For only that thought can motivate me to refrain from taking the trip unjustly if the opportunity should arise.* This is the absurdity into which Rawls is led by his rejection of "the doctrine of the purely conscientious act."[7]

Barry thus concludes that the whole idea of congruence is grounded on a mistake. It can be set aside without causing any damage to Rawls's theory as a whole. It follows that it is also a mistake for Rawls to turn to political liberalism.[8] I think Barry's criticism is unfair to Rawls. Through a careful analysis of Ross's doctrine and Rawls's view of practical reason for action in the next section, I shall show that there is indeed a need for congruence.

## 2 The Need for Congruence

The doctrine of the purely conscientious act is concerned with the nature and source of the motivational power of moral obligation. Ross claims that the goodness of morally good actions must arise from a certain kind of motive which is connected with a certain type or types of character. This kind of motivation is our sense of duty. It is stipulated as something distinct from, and superior to, other

---

[7] Barry, "John Rawls and the Search for Stability," p.889, my emphasis.

[8] In the last sentence of his long article, Barry concludes that "Rawls's sweeping recantation is uncalled-for, and that the failure of *Political Liberalism* does not discredit *A Theory of Justice*. I believe that, as time goes on, *A Theory of Justice* will stand out with increasing clarity as by far the most significant contribution to political philosophy produced in this century. Only one thing threatens to obscure that achievement: the publication of *Political Liberalism*." "John Rawls and the Search for Stability," p.915.

desires of any kind.[9] Only so will the necessity of moral obligation and the supreme worth of conscientious actions be warranted. Where does this moral motive stem from? From our pure practical reason! The recognition of an act as one's duty can by itself motivate us to act:

> There is no more mystery in the fact that the thought of an act as one's duty should arouse an impulse to do it, than in the fact that the thought of an act as pleasant, or as leading to pleasure, should arouse an impulse to do it.[10]

Ross contends that this desire to do our duty need not presuppose Kant's metaphysical view about human beings' phenomenal and noumenal nature. Rather, it simply springs from our possession of reason. As he puts it, "it is only natural that there should arise a desire, itself springing from our rational apprehension of principles of duty, not to be the slave of low desires but to regulate our life by these principles."[11] A purely conscientious act proceeds from "*a desire for a specifically distinct object*, not for the attainment of pleasure nor even for the conferring of it on others, but just for the doing of our duty."[12] Ross stresses that the word "purely" specifies the distinct nature of singleness of this desire.[13] This is essentially a Kantian account of the moral motive. A morally good action must be done from a sense of duty, which is detached from other inclinations or a

---

[9] At this point, Ross disagrees with Kant that apart from the sense of duty, there are other desires resulting from our nature as rational beings. He thus says that "we can agree with him [Kant] in thinking that the sense of duty is the highest motive, without following him in putting all other motives on the same dead level." *Foundations of Ethics* (Oxford: the Clarendon Press, 1939), p.206.

[10] W. D. Ross, *The Right and the Good* (Indianapolis: Hackett Publishing Company, 1930), pp.157-58.

[11] Ross, *Foundations of Ethics*, p.206.

[12] Ross, *The Right and the Good*, p.158, emphasis added.

[13] Ross, *Foundations of Ethics*, p.207.

person's happiness. "An action from duty is to put aside entirely the influence of inclination and with it every object of the will."[14] There is no relation between the moral motive and the good. So, a rational agent should comply with the moral law even if it infringes upon all his inclinations and interests. Thus the priority of the sense of duty is purely based on the nature of practical reason.

It should be noted that it is this particular Kantian position that Rawls describes as the doctrine of the purely conscientious act, but not Ross's more general doctrine of moral intuitionism.[15] He raises this issue not because he has a primary interest in Ross's intuitionism, but because he wants to make a comparison of this doctrine with his own view of moral motivation so that the distinctive feature of congruence and its indispensability to justice as fairness can be seen.[16]

What is wrong with Ross's account of moral motive then? Rawls offers the following explanation:

> It would seem then, that the doctrine of the purely conscientious act is *irrational*...Ross holds that the sense of right is a desire for a distinct (and unanalyzable) object, since a specific (and unanalyzable) property characterizes actions that are our duty. The other morally worthy desires, while indeed desires for things necessarily connected with what is right, are not desires

---

[14] Kant, *Groundwork of the Metaphysics of Morals* collected in *Practical Philosophy*, tran. & edited by Mary J. Gregor (Cambridge: Cambridge University Press, 1996), p.55.

[15] Rawls has clearly indicated his reference to this view in his discussion. (TJ: 477/418 rev., footnote 15)

[16] Rawls has never used the term "intuitionism" to describe Ross's theory. He refers intuitionism to such a doctrine that "there is an irreducible family of first principles which have to be weighted against one another by asking ourselves which balance, in our considered judgment, is the most just." (TJ:34/30 rev.)

> for the right as such. But on this interpretation *the sense of right lacks any apparent reason*; it resembles a preference for tea rather than coffee. Although such a preference might exist, to make it *regulative* of the basic structure of society is utterly *capricious*. (TJ:477-78/418 rev., my emphasis)

At first sight it is unclear why this doctrine is irrational and capricious. If we agree with Barry, there would be nothing wrong to hold that recognition of something as just can give rise to a sense of duty. Mendus tries to defend Rawls by saying that what Rawls specifically objects to is the concept of goodness (or rightness, or justice) as simple and unanalysable. [17] It is the rather peculiar account of moral motivation of rational intuitionism. Barry's charge is thus harmless because Rawls could agree with him on the relation between motivation and practical reason in general.[18] Nevertheless, this defence is not of much help because it cannot explain why Rawls would uphold the demanding requirement of congruence. Barry, for instance, may argue that what Rawls should do is simply to accept a Scanlonian account of the moral motive, and abandon the project of congruence. Besides, we should note that Rawls's complaint is that Ross's doctrine is irrational. He acknowledges that such a motive may exist. What he argues against is that it should be taken to be the regulative motive of the basic structure. So, more attention should be paid to the problem of rationality.

To judge whether an act is rational, we need to know Rawls's theory of rationality. Rawls holds a conception of deliberative rationality, which stipulates

---

[17] Mendus, "The Importance of Love in Rawls's *Theory of Justice*," *British Journal of Political Science*, 29, (1999), p.62.

[18] Freeman deploys the similar strategy to respond to Barry. Freeman, "Congruence and the Good of Justice," p.282.

that an act is rational when it can best achieve what an agent wants most after informed deliberation. The agent's rational desire ultimately determines his practical reason for action. This is what Parfit called a *desire-based* theory.[19] One of the distinctive features of this theory is that rationality itself cannot decide what an agent should most want. Therefore, "knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently." (PL:49, footnote 1) Regardless of this, Rawls makes a further assumption that we all live according to a plan of life. A person's rational plan of life determines his good. A plan is rational if, and only if:

> (1) it is one of the plans that is consistent with the principles of rational choice when these are applied to all the relevant features of his situation, and (2) it is that plan among those meeting this condition which would be chosen by him with full deliberative rationality, that is, with full awareness of the relevant facts and after a careful consideration of the consequences. (TJ: 408/358-359, rev.)

So, an act is rational if and only if it is the most effective way to realize what the agent most wants, namely, the plan of life that he will adopt with full deliberative rationality. It implies that a rational agent will only have a motive to act on principles that is beneficial to his plan of life. When the question "why should I be just" is raised, the answer hinges considerably on whether acting justly can be conceived as a good to the rational agent.[20] "The desire to act justly is not, then, a form of blind obedience to arbitrary principles unrelated to rational aims."

---

[19] Derek Parfit, "What We Could Rationally Will," *The Tanner Lectures on Human Value*, (Salt Lake City: Humanities Center, University of Utah, 2002), p.342, retrieved from http://www.tannerlectures.utah.edu/lectures/volume24/parfit_2002.pdf

[20] At this point, as Rawls acknowledges, his view is greatly influenced by Foot. See Foot, "Moral Belief," pp.125-30.

(TJ:476/417 rev.) To the contrary, Rawls argues:

> A theory should present a description of an ideally just state of affairs, a conception of a well-ordered society such that the aspiration to realize this state of affairs, and to maintain it in being, *answers to our good and is continuous with our natural sentiments*. (TJ:477/417 rev., emphasis added)

We may now understand why Rawls thinks that the motivational assumption of the purely conscientious act is irrational. For the doctrine fails to provide any further justification to rational individuals that they have a duty to give absolute priority to the moral sentiment over other desires. It supposes that an agent should have the highest moral motive to do what is right and just simply because it is right and just. The desire to fulfil our duty is completely detached from the desire to realize our interest. When a person asks himself why it is rational for him to fulfil his duty even if there is a strong conflict between moral command and his conception of the good, Ross's reply is that "the truest answer I can find is that I do it because, then at least, I desire to do my duty more than I desire anything else."[21] For Rawls, this answer is question-begging. It lacks any normative force to establish its motivational authority.

Justice is a virtue concerning what we owe to other people. It may require us to sacrifice our greatest interests for the sake of other's rights. Unlike other actions that would bring pain, boredom and loneliness, the desire to act justly could hardly give sufficient reasons for action by itself. As Foot remarks, " 'it is unjust' gives a reason only if the nature of justice can be shown to be such that it

---

[21] Ross, *Foundations of Ethics*, p.206.

is necessarily connected with what a man wants."[22] Foot believes that there are two types of desires for actions. The first type can directly motivate us in a certain way because it can satisfy some of our basic wants that are ultimately related to our well-being. In this case, no further justification is required. Avoiding pain is such an example.[23] On the contrary, the second type needs *further* reason to trigger our action. The sense of justice belongs to this type. Rawls shares this view with Foot:

> The desire to act justly is not a final desire like that to avoid pain, misery, or apathy, or the desire to fulfill the inclusive interest. The theory of justice supplies *other descriptions* of what the sense of justice is a desire for; and we must use these to show that a person following the thick theory of the good would indeed confirm this sentiment as regulative of his plan of life. (TJ:569/499 rev., emphasis added)

For Rawls, the "other descriptions" are mainly related to how an individual, consistent with desire-based deliberative rationality, can confirm the sense of justice as par of their good. He does not believe that rationality has such power that it is able to dictate that people should have a pure moral motive to comply with the principles of justice regardless of their conception of the good. The sense of justice is not that type of desire. If some people have it, it is irrational. His point is that a just person who lives his life from the perspective of justice can do this

---

[22] Foot, "Moral Belief," p.127.

[23] This point is famously illustrated by Hume in the following paragraph. "Ask a man *why he uses exercise*; he will answer, *because he desires to keep his health*. If you then enquire, *why he desires health*, he will readily reply, *because sickness is painful*. If you push your enquires farther, and desire a reason *why he hates pain*, it is impossible he can ever give any. This is an ultimate end, and is never referred to any other object." Hume, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L.A.Selby-Bigge (Oxford: Oxford University Press, 1975), Appendix I, V, p.293.

only if he builds the priority of justice into his conception of the good.[24] It is for this reason that Rawls finds the doctrine of the purely conscientious act irrational, and congruence necessary to stability. Rawls not only rejects Ross's specific account of the moral motive, but also disapproves of a general claim that the sense of justice detached from our good can effectively motivate rational individuals to act.

We may also think about this question from the perspective of the circumstances of justice. Rawls presumes that society is a cooperative venture for mutual advantage. All members are supposed to have a fundamental interest in advancing their conceptions of the good. Their primary motive to participate in cooperation is to realize their good. This explains why they have conflicting claims about distributive justice. Rational individuals are not expected to act from a pure sense of moral duty. To what extent they can develop an effective moral motive largely depends upon whether the principles of justice concerned can effectively promote their good. Therefore, the real problem of congruence is about "what happens if we imagine someone to give weight to his sense of justice only to the extent that it satisfies other descriptions which connect it with reasons specified by the thin theory of the good." (TJ:569/499 rev.)

Given what I have argued in this section, Rawls's project is not as unnecessary and absurd as Barry perceives. On the contrary, the quest for convergence is a necessary step to achieve moral stability. Before we go on to assess Rawls's substantive argument for congruence, an ambiguity concerning the

---

[24] I thank John Charvet for clarifying this point.

definition of the good must be settled. Rawls repeatedly reminds us that when he talks about the good of the sense of justice, the good must be understood in a thin sense. This is a puzzling reminder. We know that the main purpose of the thin theory of the good is to define primary goods in the original position. Primary goods are those things that rational individuals, whatever they are, desire as prerequisites for carrying out their plans of life. These things include rights and liberties, income and wealth, opportunity, and self-respect. Nevertheless, the sense of justice does not belong to this list. Moreover, the thin theory is mainly used in the original position. After the veil of ignorance is lifted, people will deliberate according to the full theory of the good in connection with deliberative rationality and their plans of life. Now since congruence takes place in the second stage where people have full knowledge of their situation with no restriction on information, it is unreasonable to suppose that people would still be guided by the thin theory in judging the goodness of the sense of justice. Barry, for example, has put the following question to Rawls:

> Given that the problem, as Rawls conceives it, is one of offering a reason for real people to "affirm their sense of justice," it is surely correct to specify that they should have full information. But then why should they throw away the information about their own distinctive conceptions of the good (their "thick" conception) and restrict the question to one the answer to which is going to be the same for everyone?[25]

This is a good observation, and Rawls's position is ambiguous. For example, he explicitly states that "when we come to the explanation of the social values and the stability of a conception of justice, *a wider interpretation of the good* is

---

[25] Barry, "John Rawls and the Search for Stability," pp.885-86.

required." (TJ:398/350 rev., my emphasis) He also says that the real problem of congruence is whether a rational person would adopt the standpoint of justice as a regulative good viewed "in the light of the thin theory *with no restrictions on information.*"(TJ:567/497 rev., emphasis added) If rational individuals have already known their conceptions of the good, it is implausible to expect them not to use the full theory of the good to assess their situation.

In my view, the only plausible explanation is that even though the knowledge of conceptions of the good is available to people, the thin theory should still be used to show that it is rational for *every* person to desire the sense of justice as his regulative good. It is thin in the sense that it is good for all rational plans of life. Therefore, "the plan of life which does this is his best reply to the similar plans of his associates; and *being rational for anyone, it is rational for all.*" (TJ:567/497 rev., my emphasis) The claim of rationality applies to every one. The real challenge is, given desire-based rationality and the fact of pluralism, whether Rawls can provide convincing arguments to vindicate this ambitious claim. This is what we are going to examine in the rest of this chapter.

## 3 The Free-Rider Argument

In Section 86 "The Good of the Sense of Justice", Rawls provides three major arguments to justify congruence of the right and the good. I will call them *the Free-Rider, the Social Union, and the Kantian Interpretation Arguments* respectively. These arguments are different in nature. I will examine the first argument in this section while dealing with the other two in the coming sections.

The free-rider argument is concerned with the psychological cost of being a

free-rider in a well-ordered society regulated by a public conception of justice. Rawls contends that an egoist free-rider will suffer from loss of spontaneity and naturalness if he plans a systematic course of deception and hypocrisy. He must keep on hiding himself even from those around him. Thus, the supposition that acting unjustly is more profitable than doing justice is dubious.[26]

This argument is weak. First of all, it is inconclusive. For whether a free-rider egoist would comply with principles of justice depends upon his rational calculation of cost and benefit in different circumstances. It is possible that even if an egoist knows that he has to pay some psychological cost, he will still choose to act unjustly in case free-riding will win him even greater benefits. There is nothing irrational for him to take advantage of the cooperative efforts of others. Rawls's reply is that the price of free-riding is particularly high in a well-ordered society where most other people act on, and from the sense of justice. (TJ: 570/499 rev.) But this is empirically questionable. For an egoist who does not care about the value of spontaneity and naturalness, cheating may scarcely upset him or his life. For this argument to stand, Rawls needs a much stronger account to show that spontaneity is essential to any rational plan of life, and that any act of deception would substantively hamper one's conception of the good. But Rawls cannot make such a claim about free-riders because he concedes that sometimes "acting fairly is not in general each man's best reply to the just conduct of his associates." (TJ:497/435 rev.)

Why does Rawls want to justify the good of justice to a free-rider egoist?

---

[26] Rawls here basically follows Foot's argument in "Moral Argument," p.129.

This seems unnecessary because Rawls makes it clear that his argument is not trying to show that in a well-ordered society an egoist would act from a sense of justice, nor even that he would act justly because so acting would best advance his ends. Rather, he assumes that members of that society already possess a sense of justice. What we are concerned with is "the goodness of the settled desire to take up the standpoint of justice." (TJ:568/498 rev.) However, an egoist will only work for his own interest. Even if on some occasions he acts on the principle of justice, it is merely a coincidence. An egoist will not act on moral reasons. For "having these reasons is inconsistent with being an egoist." (TJ:568/497 rev.) So, even if free-riders' self interest is congruent with justice, it is for the wrong reason.

## 4 The Social Union Argument

The social union argument rests on the Aristotelian Principle and the idea of society as a social union. The thrust is that participating in the life of a well-ordered society is a great good. And "to share fully this in this life we must acknowledge the principles of its regulative conception, and this means that we must affirm our sentiment of justice." (TJ:571/500 rev.) This argument does not address itself to egoistic free-riders by appealing to pragmatic reasons. It is grounded on the intrinsic value of community and the social nature of human kind. Rawls takes the idea of social union seriously and places much emphasis on our social nature. He aims to show that justice as fairness presupposes neither an atomistic self nor a private society. For Rawls, congruence depends in large part upon whether a well-ordered society achieves the good of community.

(TJ:520/456 rev.) [27] He explains his argument as follows:

> What binds a society's efforts into one social union is the mutual recognition and acceptance of the principles of justice; it is this general affirmation which extends the ties of identification over the whole community and permits the Aristotelian Principle to have its wider effect. Individual and group accomplishments are no longer seen as just so many separate personal goods. Whereas not to confirm our sense of justice is to limit ourselves to a narrow view. (TJ:571-72/500-501 rev.)

The Aristotelian Principle is defined as follows: "other things equal, human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity." (TJ:426/374 rev.) So, we have a natural desire to prefer exercising our higher activities and engaging in complex activities for their own sake. Rawls takes this as a basic principle of motivation to account for many of our major desires. Surprisingly, he offers little argument to prove it. He simply presumes that "complex activities are more enjoyable because they satisfy the desire for variety and novelty of experience, and leave room for feats of ingenuity and invention." (TJ:427/374 rev.)[28] This account is however insufficient to explain the complexity of human motivation. As a matter of empirical fact, people

---

[27] This argument is not taken seriously enough by critics. For example, Barry does not mention this argument at all in his "John Rawls and the Search for Stability." Although Freeman has touched upon this issue, he only interprets it as a "simplified argument from the Aristotelian principle" without discussing its connection with the idea of social union. "Congruence and the Good of Justice," pp.291-92.

[28] Rawls's another argument is that the principle is supported by an evolutionary explanation. But he immediately acknowledges that "the evolutionary explanation, even if it is correct, is not of course a justification for this aspect of our nature. In fact, the question of justification does not arise." (TJ:432/379 rev.) It is not easy to make sense of this statement. For without sufficient support, how can Rawls claim that this principle of motivation is a "natural fact", and explain many of our reasons for action?

may not choose to engage in a more complex activity because of the limit of their talents, the training cost, the toughness of the activity, and the recognition of the value of this activity by their associates. Many factors affect a rational agent's choice. Rawls actually admits this possibility. He agrees that a man whose only good is counting blades of grass should not be regarded as irrational if this is a choice that he would choose with deliberative rationality and regard as a final end regulating the schedule of his actions. As a result, "the correctness of the definition of a person's good in terms of the rational plan of life for him does not require the truth of the Aristotelian Principle." (TJ:433/380 rev.) If this is the case, why would Rawls still appeal to this principle?

Rawls's contention is that the Aristotelian Principle plays an important role in accounting for our considered judgments of value. It is then rational for persons to realize and preserve those complex capabilities, and give them a prominent place in their plans of life. The Aristotelian Principle is part of the background to the specification of our good and explains our desire for the exercise of certain capacities. If this argument succeeds, Rawls can then put forth a further claim that the exercise of the sense of justice is generally experienced as a good. As Freeman remarks, "this capacity admits of complex development and refinement. Since all have a sense of justice in a well-ordered society, it is rational for each to develop it as part of his or her plan of life."[29] The principle will then provide direct support for congruence because we have a natural desire to realize our moral capacity.

However, this argument is implausible. First, as demonstrated above, there is

---

[29] Freeman, "Congruence and the Good of Justice," p.291.

no conclusive argument to support the claim that a rational plan of life must meet the Aristotelian Principle. The argument at most describes a natural tendency of human motivation, but not an invariable pattern of choice. Second, human beings have different kinds of mature capacities. Which capacity they choose to exercise depends upon their conceptions of the good. The Aristotelian Principle itself cannot account for why rational individuals should have a desire to develop their sense of justice. Rawls admits this difficulty because he agrees that "by itself the principle simply asserts a propensity to ascend whatever chains are chosen. It does not entail that a rational plan includes any particular aims, nor does it imply any special form of society." (TJ:430/377-378 rev.) Thus the principle falls short of affirming the central place of the sense of justice in people's rational plans of life. Finally, even if people have an interest in exercising the sense of justice, it does not entail that they will give priority to the sense of justice over other capacities. As Freeman rightly asks, "what is to prevent my giving weight to my sense of justice only according to its relative intensity and subordinating it to stronger dispositions, weighting off my concern for justice against other final ends in ordinary ways?"[30] The Aristotelian Principle itself is insufficient to provide support for congruence. Rawls needs a stronger argument to show that exercising the sense of justice is not just a good, but a *common regulative good* embedded in every rational plan of life. This is indeed what Rawls intends to offer. The Aristotelian Principle just sets the background for the idea of the well-ordered society as a form of social union.

Rawls starts his argument with a comparison between the notion of private

---

[30] Freeman, "Congruence and the Good of Justice," p.292.

society and of social union. The idea of private society has three major features. First, people have competing or independent private ends which are not complementary to each other. Second, people are not moved by a desire to act justly. They are egoistic beings. What they care about is how to gain the largest share of resources through the most efficient arrangement. Finally, given the first two features, people view institutions as a means or even a burden to realize their interests. Social cooperative schemes are not thought to have value in themselves. A Hegelian civil society or a competitive market is a paradigm description of such a society. Obviously, a private society has no capacity for moral stability. Its maintenance of social order will have to heavily rely on coercive power. Rawls claims that a well-ordered society is not a private society because its members are presumed to have an effective sense of justice. More importantly, thanks to the sociability of human beings, a well-ordered society is a form of social union:

> The social nature of mankind is best seen by contrast with the conception of private society. Thus human beings have in fact *shared final ends* and they value their common institutions and activities as *good in themselves*. We need one another as partners in ways of life that are engaged in for their own sake, and the successes and enjoyments of others are necessary for and complementary to our own good. (TJ:522-23/458 rev., my emphasis)

The idea of society as a social union consists of three features. (1), it has a shared final end guiding members' actions; (2), social institutions are regarded as good in themselves; and (3) all participants find great satisfaction in the realization of shared ends. When a society possesses these features, it is a social union. The idea of social union is grounded on a rather self-evident fact about human beings: no one can fully realize his potentialities and capabilities as he

wishes because of natural and social constraints. Each individual at a time can only participate in and achieve perfection of a particular activity. The complexity of activities inevitably limits the expression of powers of an individual. Therefore, we cannot but make choices about what kinds of ability and interests we want to pursue. Nevertheless, we need not regret the incompleteness. This is because "through social union founded upon the needs and potentialities of its members that each person can participate in the total sum of the realized natural assets of the others." (TJ:523/459 rev.) By mutual cooperation, members enjoy one another's perfections of different kinds of talents. We may call this a tacit division of labour for a common project. Members develop their faculties in activities that they have chosen so as to realize the rich powers of all in their joint performance. Their ends are not competing and independent of one another. Rather, they are often complementary. Living in a social union enriches our life and increases our satisfaction.

It is surely right that the complementarity of the good is important to everyone. What is unclear is how it can lend support to the congruence argument. For instance, a free-riding egoist can appreciate the good resulting from the realization of the potentialities of the others. Of course, an egoist's end is complete and independent. But he can recognize that his realization of his final end may involve instrumental dependence on the ends of others so that the attainment of his end cannot be separated from the ends of others.[31] The question of whether a rational individual can be motivated to act justly is different from the question of whether we need one another to realize a totality of human capacities.

---

[31] I thank John Charvet for pointing out this to me.

The latter has no bearing on the former. An instrumentalist view of society seems compatible with the notion of social union. Rawls denies this in holding that only in a social union will people have a common end that informs their action. Rawls uses games as an analogy to demonstrate this salient point. He says:

> [W]e can easily distinguish four sorts of ends: the aim of the game as defined by its rules, say to score the most runs; the various motives of the players in playing the game, the excitement they get from it, the desire for exercise, and so on, which may be different for each person; the social purposes served by the game which may be unintended and unknown to the players, or even to anyone in the society, these being matters for the reflective observers to ascertain; and then finally, *the shared end, the common desire of all the players that there should be a good play of the game. This shared end can be realized only if the game is played fairly according to the rules, if the sides are more or less evenly matched, and if the players all sense that they are playing well.* But when this aim is attained, everyone takes pleasure and satisfaction in the very same thing. A good play of the game is, so to speak, a collective achievement requiring the cooperation of all. (TJ:525-26/460-61 rev., my emphasis)

Rawls holds that, like games, social unions have their shared ends and members will have a regulative and effective desire to realize them. I do not think this argument can support congruence. First of all, given that there are different sorts of ends in a game, it is dubious how the priority of the shared ends over others is warranted. Many games are competitive by nature. Winning the game and gaining the reward are often players' strongest motive to take part in a competition. Of course, this desire is not necessarily incompatible with the desire to play a good game. Even if one loses, he may still find great enjoyment in jointly performing a good game. However, in case two ends come into conflict, it

cannot be said that the common end will prevail by default. Rawls may reply that the pleasure drawn from the shared end far exceeds that from other ends. The validity of this claim, however, can only be judged on a case by case basis. Even in the same game players may have different levels of enjoyment. It is simply not the case that when the shared end is realized, all participants will find satisfaction in the very same thing with equal degrees.

Furthermore, the priority of the shared end cannot be justified by claiming that the shared end makes members' good complementary to one another and thus brings a great deal of pleasure to each participant. This is essentially an *individualistic* argument. For the appeal of social union is largely ascribed to its instrumental benefit to individuals' separate ends. Before they join a union, members are supposed to have their distinct plans of life. Rawls does not describe the shared end as an independent social entity imposed on members. For that will violate the principle of the separateness of persons. To what extent the shared end is honoured relies on members' *subjective* endorsement. As Rawls himself acknowledges, "whether individuals have a shared end depends upon the more detailed features of the activity to which their *interests* incline them as these are regulated by principles of justice." (TJ:526/461 rev., my emphasis) In this case, the notion of social union is a weak conception of community whose value ultimately depends on its contribution to individuals' pre-given conceptions of the good. The social union is a form of association that people can freely join for mutual benefit through cooperation. So, like other ends, the motivational force of the shared end is the outcome, rather than the precondition, of our choice.[32]

---

[32] Note that I am not saying that the shared end is determined by our choice. This seems too good

This weak conception of community is sharply different from a constitutive conception according to which "community describes not just what they *have* as fellow citizens but also what they *are*, not a relationship they choose (as in a voluntary association) but as attachment they discover, not merely an attribute but a constituent of their identity."[33] Rawls denies this communitarian ideal of community. One reason is that only the weak conception is consistent with desire-based rationality and the idea of the free person with a moral capacity for a conception of the good. A constitutive conception of community would limit our freedom to form, to revise, and to pursue our conceptions of the good. Thus, regardless of his emphasis on the value of social union, Rawls would not want to adopt a constitutive conception of community.

My discussion has so far focused on the general account of social union and has shown its insufficiency to prioritize the shared end over others. We now turn to examine the idea of a well-ordered society as a social union of social unions. According to Rawls, a well-ordered society manifests two essential features of social union, namely, "the successful carrying out of just institutions is the shared final end of all the members of society, and these institutional forms are prized as good in themselves."(TJ:527/462 rev.) Why these two features? Rawls's answer explains as follows:

> In much the same way that players have the shared end to execute a good and fair play of the game, so the members of a

---

to be true. It is possible that there is a final shared end of a social union even though a number of members do not have any interest in it. Since our concern is the motivating force of the shared end, we need not worry about this possibility here.

[33] Michael Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982), p.150.

> well-ordered society have the common aim of cooperating together to realize *their own and another's nature* in ways allowed by the principles of justice... and when everyone acts justly, all find satisfaction in the very same thing. (TJ:527/462 rev., emphasis added)

Participating in social cooperation matters because it is an effective means to realize each member's nature, which in turn brings them great satisfaction. Our common interest is to express our nature. To make this argument succeed, two questions must be answered: first, why should our nature be understood in this way? Second, why should the successful implementation of a just basic structure be necessary condition of realising our nature? Rawls's answer is grounded on what he calls the Kantian interpretation of justice as fairness. Whether this interpretation is well justified will be scrutinized in the next section. However, if the value of social union depends on a higher-order interest in realising our nature, then the idea of social union itself would appear not to have independent justificatory force.

This does not mean, however, that the well-ordered society is only instrumental in satisfying our independent ends. For Rawls insists that when the idea of social union is applied to the basic structure as a whole, social institutions will be regarded as good in themselves by members. Does this imply that institutions are still valuable in themselves even if a majority of members disrespects them for whatever reason? This does not make any sense. What Rawls means must be that living in a just society and doing what justice requires are *necessary* conditions of the realization of our nature. Since this is our final end, the good of well-ordered society will not serve any further ends. If so, the real force of congruence comes primarily from the Kantian interpretation of human

nature rather than any distinct feature of social union.

In this and the previous sections, I have discussed the free-rider and the social union arguments for congruence. The free-rider argument fails because we can never provide sufficient reasons to convince an egoist that being a just person is always good for him. The social union argument is not sound because neither the Aristotelian Principle nor the idea of social union are sufficient to show the good of granting priority to the sense of justice. When the idea of social union is applied to the well-ordered society, it is clear that its shared final end is based on a Kantian conception of human nature. Since the Kantian interpretation is the most fundamental argument for stability, the rest of this chapter will be devoted to it.

## 5 The Kantian Interpretation and Liberal Perfectionism

The central idea of the Kantian interpretation of justice as fairness is that rational persons would have a regulative desire to express their nature as free and equal rational being. The idea is primarily derived from Kant's conception of the moral person and the notion of autonomy. Rawls holds that justice as fairness can be interpreted as the most adequate expression of this Kantian ideal. It provides the strongest reason for rational agents to uphold justice because acting justly is itself a regulative good judging from the rational point of view. In the following discussion, I will first explicate this interpretation and then argue that it has presupposed a form of perfectionism.

It would be instructive to begin with an examination of how the right and the good are congruent with each other under the Kantian interpretation of human nature. According to Rawls:

> Acting justly is something we want to do as free and equal rational beings. The desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire. When someone has true beliefs and a correct understanding of the theory of justice, these two desires move him in the same way. They are both dispositions to act from precisely the same principles: namely, those that would be chosen in the original position. (TJ:572/501 rev.)

Having explained the possible convergence of two desires, Rawls continues to argue for its regulative priority:

> The desire to express our nature as a free and equal rational being can be fulfilled only by acting on the principles of right and justice as having first priority. This is a consequence of the condition of finality: since these principles are regulative, the desire to act upon them is satisfied only to the extent that it is likewise regulative with respect to other desires. It is acting from this precedence that expresses our freedom from contingency and happenstance. Therefore, *in order to realize our nature we have no alternative but to plan to preserve our sense of justice as governing our other aims.* This sentiment cannot be fulfilled if it is compromised and balanced against other ends as but one desire among the rest. (TJ:574/503 rev., emphasis added)

Rawls's argument can be formulated as having the following thirteen steps.

> (1) According to the Kantian account of human nature, we are essentially free and equal rational beings.
>
> (2) Rational beings have a fundamental desire to express their nature.
>
> (3) The realization of one's nature is a supreme good for a rational person.

(4) The necessary and sufficient condition of realising one's nature is to act on principles of justice which are chosen under conditions that fully represent one's nature as free and equal rational being.[34] This is because "to express one's nature as a being of a particular kind is to act on the principles that would be chosen if this nature were the decisive determining element." (TJ:253/222 rev.)

(5) The original position is designed to fulfil the task in (4). As a device of representation, it specifies conditions in which principles of justice are chosen by rational free persons in an initial situation of equality without being affected by their social position or natural endowments.

(6) Since Rawls's principles of justice would be chosen by free and equal parties in the original position, acting from these principles is therefore the most effective way to express their nature.

(7) The desire to apply and to act from the principles of justice is called a sense of justice. (TJ:567/496-497 rev.)

(8) Taking the above premises together, it can be concluded that "acting justly is something we want to do as free and equal rational beings. The desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire." (TJ: 572/501 rev.)[35]

(9) Thus it is rational for an individual to affirm his sense of justice because the realization of his nature is a supreme good

---

[34] Rawls uses "to express one's nature" and "to realize one's nature" interchangeably to refer to the same meaning.

[35] Similarly, after rejecting the doctrine of the purely conscientious act as irrational, Rawls asserts that "for one who understands and accepts the contract doctrine, the sentiment of justice is not a different desire from that to act on principles that rational individuals would consent to in an initial situation which gives everyone equal consideration as a moral person. Nor is it different from wanting to act in accordance with principles that express men's nature as free and equal rational beings." (TJ:478/418 rev.)

in virtue of (3).[36]

(10) Furthermore, the condition of finality requires that the principles of justice chosen in the original position must be regulative and overriding. The primacy of justice is assured by this formal condition.

(11) Since the principles of justice are regulative, the desire to act upon them must also be regulative. If the sense of justice is compromised and balanced against other ends as but one desire among the rest, we fail to fully realize our nature. For "how far we succeed in expressing our nature depends upon how consistently we act from our sense of justice as *finally regulative*." (TJ:575/503 rev., emphasis added)

(12) Therefore, it is always rational for individuals in a Rawlsian well-ordered society to give first priority to the sense of justice as a result of congruence.

(13) Finally, congruence is verified and justice as fairness is shown to be the most stable conception of justice.

We can note that the whole argument is grounded on a particular interpretation of human nature. If the argument holds true, justice as fairness will be as stable as one can hope for. There will be no disharmony between the right and the good. A well-ordered society regulated by justice as fairness is not only reasonable from the impartial perspective of the original position, but also desirable from the first-person rational point of view. If one is rational enough, he will recognize that being a just person is exactly what he most desires to be. The congruence argument has a deep appeal because it offers an answer to the

---

[36] Freeman holds that this is a result of the Aristotelian Principle. Nevertheless, according to Rawls, our good is determined by the satisfaction of our rational desires. Since the expression of our nature is a rational desire, it is not necessary to appeal to the Aristotelian Principle to affirm its value. Freeman, "Congruence and the Good of Justice," pp.293-94.

question of why a rational person would have sufficient reason to act morally.

It should be noted that this Kantian conception of human nature points to a perfectionist position. According to Rawls, perfectionism is a type of ideal-regarding principle which directs society to arrange institutions and to define the duties and obligations of individuals so as to realize an ideal independently specified. (TJ:325-26/285-286 rev.)[37] Thomas Hurka, a prominent perfectionist philosopher, also defines perfectionism as such a doctrine that "the good consists at bottom in developing one's 'nature', or realizing a 'true self': certain properties are central to one's identity, and one's good consists in developing these properties to a high degree."[38] Furthermore, perfectionism is based on an objective theory of the good. The realization of good in a life makes that life better independently of how much it is wanted or enjoyed.

The Kantian interpretation of justice as fairness seems to fit the definition of perfectionism perfectly. It prescribes an objective conception of human nature as free and equal rational being; it holds that a justifiable conception of justice must be able to express our nature as fully as possible; it even stipulates that acting in accordance with justice is a constitutive good with the highest priority. At some point, Rawls accepts this characterisation of his theory. For instance, he remarks that "a certain ideal is embedded in the principles of justice, and the fulfilment of

---

[37] Rawls holds that there are two variants of perfectionism. In the first, it is the sole principle of a teleological theory which aims to maximize the achievement of human excellence; in the second, the principle of perfection is only one standard among several in an intuitionist theory. (TJ:325/285-286 rev.) This indicates that maximization is not a defining feature of perfectionism. What is crucial is the role of human excellence played in a theory. I will elaborate this point in more detail later.

[38] Thomas Hurka, "Perfectionism" in E. Craig ed., *Routledge Encyclopedia of Philosophy,* (London: Routledge, 1998), retrieved from http://www.rep.routledge.com/article/L070SECT5.

desires incompatible with these principles has no value at all. Moreover we are to encourage certain traits of character, especially a sense of justice." (TJ:326-27/287 rev.) Despite this, Rawls insists that "perfectionism is denied as a political principle." (TJ:329/289 rev.) Justice as fairness at most occupies "an *intermediate position* between perfectionism and utilitarianism." (TJ:327/287 rev., emphasis added)

Rawls offers two arguments to defend his position. First, he stresses that his ideal conception of moral personality does not invoke "a prior standard of human excellence." (TJ:327/287 rev.) This explanation is dubious. If justice as fairness is founded on a prior conception of human nature, it will entail a prior standard of excellence. The standard is to realize our nature as free and equal rational being as fully as possible. It determines how the original position would be designed and what kind of principles would be chosen; rational citizens are also expected to have a higher-order interest to develop their moral capacity for a sense of justice and for a conception of the good. "For this sentiment reveals what the person is." (TJ:575/503 rev.) In this regard Rawls is undeniably a liberal perfectionist.

Rawls's second argument is that any principle of perfection would be rejected by contractors in the original position because it fails to provide a firm basis for the equal liberties in a pluralistic society. Although contractors are not cognizant of their particular moral and religious interests, they are aware that they will be devoted to different conceptions of the good in a well-ordered society. Any perfectionist principle will be incompatible with equal basic liberties for all. So, they would not "risk their freedom by authorizing a standard of value to define what is to be maximized by a teleological principle of justice." (TJ:328/288-289

rev.)

The validity of this argument is based on an assumption that the principle of equal liberty itself does not depend upon any perfectionist ideal. This assumption is disputable though. It should be recalled that the reason for the contractors to give top priority to liberty is mainly that they regard themselves as free and equal rational beings. For Rawls, the principle of equal liberty is the only alternative compatible with our nature, and "to express one's nature as a being of a particular kind is to act on the principles that would be chosen if *this nature were the decisive determining element.*" (TJ:253/222 rev., emphasis added) The original position serves as a mediating idea to represent our nature as free and equal. Our differences in social class and natural endowment are excluded from the original position because they would vitiate our nature as autonomous equal beings. To act on principles derived from these factors is to act heteronomously. (TJ:252/222 rev.) Similarly, the fact that we affirm a particular ideal does not give us a good reason to expect others to accept a conception of justice in that ideal's favour because it is incompatible with our capacity to form, revise, and pursue a conception of the good. This explains why the contractors are to be ignorant of their philosophical and religious worldviews. All the above clearly shows that in setting up the original position, the Kantian ideal has already been incorporated in its description. It plays a *decisive determining role* in deriving Rawls's principles of justice. In this sense what the original position represents is not a *neutral* point of view. Rather, it embodies a distinctive liberal conception of the person. "The parties conceive of themselves as free persons who can revise and alter their final ends

and who give priority to preserving their liberty in this respect." (TJ:475, rev.)[39] My claim that Rawls is a liberal perfectionist is thereby sustained.

That being said, it may be argued that Rawls's perfectionism is a second-order ideal that can accommodate a variety of *first-order* substantive conceptions of the good. The Kantian ideal encourages people to develop their capacity as autonomous rational agents through choosing their own ways of life. But it will not privilege any particular conception of the good at the ground-floor level. As Barry puts it, "anything could be regarded as good (in a second-order way) so long as the person who conceived it as good (in a first-order way) had arrived at this conception in a way that satisfied the requirements of autonomy."[40]

This defence, if valid, seems to have the advantages of two worlds. On the one hand, it can admit that liberalism is grounded on a particular interpretation of human nature; on the other hand, it can avoid direct competition with other conceptions of the good by positing itself in a higher-order position. Many liberals hold this position. Kymlicka, for example, believes that this is the moral foundation of Rawls's theory. Moreover, the value of personal autonomy does not lead to perfectionism. In contrast, state neutrality is required to respect people's self-determination.[41] I am not convinced by this account. Nor do I believe that this is a proper interpretation of Rawls. In the rest of this section, I will argue that

---

[39] The most fundamental change in the revised edition of *A Theory of Justice* is the re-definition of the parties in the original position as having a common higher interest in developing their two moral powers -- their capacity for a sense of justice and their capacity for a conception of the good -- in order to secure the priority of basic liberty. This indeed confirms my claim that the derivation of the principle of liberty depends upon a perfectionist account of human nature and the corresponding human interest. See TJ: xii, rev.

[40] Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), p.129.

[41] Kymlicka, *Contemporary Political Philosophy* (Oxford: Clarendon Press, 1990), p.207.

even Kymlica himself has adopted a theory of liberal perfectionism.

According to Kymlicka, the starting point of liberalism is that we all have an essential interest in leading a good life. However, leading a good life differs from leading the life we currently believe to be good. For our belief about value could be mistaken. We may misunderstand our real interest or misjudge the value of a particular activity. It follows that we should be able to stand back from our existing ends and deliberate whether our plan of life is really worth pursuing. This does not mean that one who believes that he is in a better position to know what is good can impose his view on another person because "no life goes better by being led from the outside according to values the person does not endorse."[42] This is what Kymlicka calls "endorsement constraint." No matter how good a way of life may be from a third-person perspective, it cannot make a person's life better if it is not accepted by that person from inside, according to his beliefs about value. In Kymlicka's view, the endorsement constraint is applicable to most valuable forms of human activity.[43] As a consequence, leading a good life requires two pre-conditions:

> One is that we lead our life from the inside, in accordance with our beliefs about what gives values to life; the other is that we be free to question those beliefs, to examine them in the light of whatever information, examples and arguments our culture can

---

[42] Kymlicka, *Contemporary Political Philosophy*, p.203.

[43] Kymlicka, however, admits that sometimes short-term state intervention is justifiable if we accept that "one way to get people to pursue something for the right reasons is to get them to pursue it for the wrong reasons, and hope that they will then see its true value." Therefore, "the endorsement constraint argument, by itself, cannot rule out all forms of state perfectionism." Once this qualification is granted, the liberal objection to perfectionism is no longer as strong as it primarily claims. *Contemporary Political Philosophy*, p.233.

provide.[44]

These conditions justify the value of self-determination. Kymlicka then claims that this account manifests Rawls's conception of the free person. A free person is characterized as capable of forming, revising, and rationally pursuing a conception of the good.[45] In a nutshell, free persons "think of themselves not as inevitably tied to the pursuit of the particular final ends they have at any given time, but rather as capable of revising and challenging these ends on reasonable and rational grounds."[46] Rawls further stipulates that free persons are moved by a higher-order interest to exercise this distinctive power of self-determination. It is higher-order in a sense that it is supremely regulative and effective. It governs our deliberation and conduct whenever circumstances are relevant to its fulfilment.[47] Why should we have such a regulative desire to preserve this capacity? Kymlicka argues that Rawls's answer must be that we have an essential interest in leading a good life.[48] Nevertheless, the commitment to self-determination does not lead to perfectionism. On the contrary:

> Rawls argues that this account of self-determination should lead us to endorse a 'neutral state'—i.e. a state which does not justify its actions on the basis of the intrinsic superiority or inferiority of conceptions of the good life, and which does not deliberately attempt to influence people's judgments of the value of these different conceptions.[49]

---

[44] Kymlicka, *Liberalism, Community and Culture* (Oxford: Clarendon Press, 1990), p.13.

[45] Another feature is our capacity for a sense of justice, the capacity to understand, to apply, and to act from the public conception of justice. (PL:19)

[46] Rawls, "Kantian Constructivism in Moral Theory" in *Collected Papers*, ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), p.309.

[47] Rawls, "Kantian Constructivism in Moral Theory," p.312.

[48] Kymlicka, *Liberalism, Community and Culture*, p.12.

[49] Kymlicka, *Contemporary Political Philosophy*, p.205.

The connection between self-determination and neutrality is thereby established. Kymlicka calls this conception of neutrality *justificatory neutrality*.[50] This claim of neutrality, I believe, is misleading. By definition, a conception of the good normally consists of a more or less systematic account of what is valuable in human life. It provides a framework through which we can rank our preferences and give meaning to our life. It guides our action. According to Rawls, Kant and Mill's liberalism are examples of comprehensive conceptions of the good because they appeal to autonomy and individuality respectively to inform our thought and conduct as a whole. (PL: 37, 78)

In view of this, it is fair to say that Kymlicka's self-determination-based liberalism has presupposed a conception of the good. For instance, he has repeatedly borrowed support from Mill to vindicate the importance of autonomy in leading a good life.[51] He also objects to Rawls's political liberalism by arguing that the value of autonomy should not only be limited to the political sphere. Rather, it should govern human thought and action generally.[52] Kymlicka's message is clear: forms of life are truly valuable for us only if we perceive them as ones we endorse, or would endorse in a reflective and critical manner. To lead a good life, we must regard ourselves as autonomous beings who can freely and rationally question our beliefs. A liberal should deem exercising our capacity for self-determination a regulative interest, and respect for independence and individuality. Although autonomy allows a wide range of choice of different

---

[50] Another conception of neutrality is called "consequential neutrality" which requires that the state should seek to help or hinder different life-plans to an equal degree. Kymlicka, "Liberal Individualism and Liberal Neutrality" in *Communitarianism and Individualism* ed. Shlomo Avineri and Avner de-Shalit (New York: Oxford University Press, 1992), p.166.

[51] Kymlicka, *Liberalism, Community and Culture*, pp.9-19.

[52] Kymlicka, *Multicultural Citizenship* (Oxford: Oxford University Press, 1995), p.160.

substantive views of the good life, it shapes our life in a fundamental way.

This liberal ideal is an object of reasonable disagreement though. It is not even widely shared in Western democratic societies. Kymlicka admits that there are many existing non-liberal minority groups who do not give priority to self-determination over their religious belief and cultural practice. The later Rawls is also well aware that many people living in a democratic society may not value autonomy at all:

> They may have, and often do have at any given time, affections ,devotions, and loyalties that they believe they would not, and indeed could and should not, stand apart from and evaluate objectively. They may regard it as simply unthinkable to view themselves apart from certain religious, philosophical and moral convictions, or from certain enduring attachments and loyalties. (PL:31)

This indicates that in Rawls's mind, self-determination-based liberalism embodies a perfectionist conception of the good. Kymlicka is thus wrong to claim that self-determination requires state neutrality. Kymlicka counters that even though a liberal society encourages rational assessment and revisions of one's ends, it does not compel people to lead a particular form of life. Hence, "even if this view of autonomy conflicts with a religious minority's self-understanding, there is no cost to accepting it for political purpose."[53] Nevertheless, this is merely Kymlicka's wishful thinking. The court case *Winsonsin vs. Yoder* which Kymlicka cites is a good example. The Amish community requests the government to allow them to withdraw their children from school before the age

---

[53] Kymlicka, *Multicultural Citizenship*, p.160.

of 16 in accordance with its religious doctrine. Kymlicka argues that this kind of internal restriction is unacceptable because it violates children's personal autonomy. But from the Amish point of view, they are paying a heavy cost for living in a liberal society. Liberal autonomy is not as neutral as Kymlicka claims. As Barry rightly points out, this second-order conception of the good as autonomy actually requires that "only those conceptions that have the right origins—those that have come about in ways that meet the criteria for self-determined belief—can form a basis for activity that has value."[54]

We can now conclude that if Kymlicka's interpretation of justice as fairness is right, then Rawls is a perfectionist instead of a neutralist. [55] However, this second-order conception has great difficulty in justifying congruence of the right and the good. This difficulty involves two incompatible conceptions of freedom stemming from Kant's moral philosophy. The following section will focus on this issue.

## 6 Neutral Freedom and Good Freedom

The notion of self-determination stipulates that we should be free to deliberate and to choose our ways of life. It manifests a conception of personality as free agency. According to Frankfurt, the very concept of a person consists in having desires of the second order about first-order desires. Unlike animals, persons can form the second-order volitions and make choices according to their

---

[54] Barry, *Justice as Impartiality*, p.132.

[55] Rawls expressly approves Kymlicka's interpretation as "on the whole satisfactory" although "modulo adjustments that may need to be made to fit it within political liberalism as opposed to liberalism as a contemporary doctrine."(PL:27) This confirms my claim that both early Rawls and Kymlicka have adopted a comprehensive liberal conception of the good.

will.[56] A free agent has the capacity for reflective self-evaluation of his desires and beliefs. In view of this, freedom of choice enables us to manifest our identity as free agent. Rawls endorses this conception of agency when he stresses that a free person is able to form, examine, and revise his conception of the good. Since persons have the capacity to make free and rational choice, they are presumed to be responsible for their ends.[57]

This conception of agency seems to imply that our freedom is equally manifested in choosing between good and evil, as much as in choosing different conceptions of the good. The upshot is that we choose in accordance with our second-order volitions. Self-determination itself does not prescribe what we should choose. We cannot conclude that a good voluntary action expresses a greater degree of freedom than an evil one. Sidgwick calls this "*Neutral Freedom*"—"freedom exhibited in choosing wrong as much as in choosing right."[58]

This conception of freedom poses a serious challenge to Rawls's congruence project: If congruence depends upon a person's higher-order interest in expressing his nature as free being, and if the exercise of freedom is neutral between right and wrong, how can the desire to express freedom necessarily move the rational person to honour the regulative priority of justice? Does he not equally realize his nature when he chooses to act unjustly after informed deliberation? As Sidgwick

---

[56] Harry Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), pp.11-25.

[57] Rawls, "Social Unity and Primary Goods," in *Collected Papers*, pp.369-70.

[58] Sidgwick, *The Methods of Ethics* (Indianapolis: Hackett Publishing Co., 1981), seventh edition, p.513.

famously puts it, "the scoundrel must exhibit and express his characteristic self-hood in his transcendental choice of a bad life, as much as the saint does in his transcendental choice of a good one."[59] Following the neutral conception, freedom of choice and compliance with justice as fairness do not appear to necessitate each other. Nor can Rawls say that people who act unjustly are unfree. In short, the commitment to neutral freedom (or self-determination) does not result in congruence.

One possible way to resolve this difficulty is to appeal to a more "positive" conception of freedom. This conception must be able to provide resources for Rawls to say that abiding by principles of justice is the *only* way to realize our nature as a free person. A person is free in proportion as he realizes his nature. Therefore, although people have neutral freedom to make their choice, they realize their *true freedom* by acting on moral principles which express that designated end. This conception is what Sidgwick calls "*Good*" or "*Rational Freedom*", according to which freedom consists in one's obedience to rationality, or moral laws based on pure practical reason.[60] A person is heteronomous if he is moved to act by his non-rational desires.

When Sidgwick draws this distinction between neutral and good freedom, he is talking about Kant's moral philosophy. He suggests that both conceptions can be found in Kant. When Kant has to connect the notion of freedom with that of moral responsibility and free will, he refers to neutral freedom. When he intends to prove the possibility of unconditional obedience to moral law as such without

---

[59] Sidgwick, *The Methods of Ethics*, p.516.

[60] Sidgwick, *The Methods of Ethics*, p.512.

the intervention of sensible impulses, and to exhibit the independence of reason in influencing choices, he refers to Good Freedom.[61] Sidgwick points out that these two conceptions are incompatible with each other. One cannot be described as free in making choices while being unfree in making wrong decision. The conditions of exercising neutral freedom and good freedom are entirely different. To avoid this paradox, Sidgwick argues that Kant must drop either of them to make his use of freedom consistent.

Rawls is well aware of Sidgwick's critique of Kant. (TJ:254-56/224-225 rev.) Surprisingly, I find him following Kant in employing both conceptions of freedom in his theory. We have already seen that in justifying the priority of the principle of liberty and responsibility for our choice, Rawls's conception of the free person is defined by neutral freedom. We are free when we choose our ends. But in his response to Sidgwick's criticism against Kant, Rawls turns to the conception of good freedom. He says:

> Kant's reply [to Sidgwick] must be that though acting on any consistent set of principles could be the outcome of a decision on the part of the noumenal self, not all such action by the phenomenal self expresses this decision as that of a free and equal rational being. Thus if a person realizes his true self by expressing it in his action, and he desires above all else to realize this self, then he will choose to act from principles that manifest his nature as a free and equal rational being. (TJ:255/224 rev.)

Rawls does not deny that both conceptions of freedom appear in Kant's theory. However, he stresses that although we can freely choose and act on any

---

[61] Sidgwick, *The Methods of Ethics*, p.513.

consistent principles, these principles do not equally express our true self as free and equal beings. Only when we act from those that manifest our nature can we become truly free. Since our nature is independently defined, this conception of freedom must refer to good freedom. For Rawls, his principles of justice are the most adequate option consistent with our nature because it is the result of agreement in the original position:

> The parties qua noumenal selves have *complete freedom* to choose whatever principles they wish; but they also have a desire to express their nature as rational and equal members of the intelligible realm with precisely this liberty to choose, that is, as beings who can look at the world in this way and express this perspective in their life as members of society...Thus men exhibit their *freedom*, their independence from the contingencies of nature and society, by acting in ways they would acknowledge in the original position. (TJ:255-56/225 rev., emphasis added)

Both conceptions of freedom are mentioned in the above paragraph. Rawls's main argument can be summarized as follows.

> 1. We are by nature free and equal rational beings.
>
> 2. We have a regulative desire to realize our nature.
>
> 3. The parties of the original position have complete *freedom* to choose whatever principles of justice they wish. (Neutral Freedom)
>
> 4. They would choose Rawls's principles because they could most fully express their nature as *free* beings. (Good Freedom)
>
> 5. Therefore, acting unjustly is acting in a manner that fails to exhibit our true self as free being in the well-ordered society. In

this sense, an unjust person is an unfree person.

To make this argument valid, Rawls needs to demonstrate that the parties in the original position must have a common desire to choose Rawls's principles to express their *good* freedom even though they have complete *neutral* freedom to choose. The crucial question is whether both conceptions of freedom can coherently exist in this argument. First of all, if we adopt the conception of neutral freedom, there seems no necessary connexion between freedom of choice and the derivation of Rawls's principles. The very idea of free choice implies that people can make different choices for different reasons. Even in the highly restricted original position, as some critics have pointed out, the parties may have reason to choose the principle of average utility.[62] Of course Rawls would argue that the acceptance of his principles "is the *only* choice consistent with the full description of the original position. The argument aims eventually to be strictly *deductive*." (TJ:121/104 rev., my emphasis) But if so, the idea of contract becomes redundant and the claim that people have freedom to choose is a fraud. By the same token, there is no ground for Rawls to claim that people are unfree if they decide to act on non-Rawlsian principles after the veil is lifted. Therefore, (4) and (5) cannot be inferred from (3).

If we look at Rawls's argument more closely, we will notice that the conception of neutral freedom plays no substantive role in the argument. Even if we drop (3) , (4) and (5) can still be derived from (1) and (2). The whole argument

---

[62] For example, John Harsanyi, "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," in *John Rawls: Critical Assessment* vol.1, ed. Chandran Kukathas (London & New York: Routledge, 2003), pp.216-38. Also see my discussion in the previous chapter.

is actually based on the conception of good freedom. Rawls's main idea can be put as follows: since we have a higher-order desire to express our nature as free being, we will only act on a conception of justice that can manifest our good freedom most fully. Rawls's principles are thus deductively derived. If there is any difficulty in reaching this conclusion, Rawls could simply modify the description so that the parties would "choose" his principles unanimously.

We now see that the two conceptions of freedom cannot be used consistently in the same argument. Rawls faces the same critique as Sidgwick directed against Kant. In particular, in his argument for congruence, Rawls must adopt the conception of good freedom. For if people manifest their freedom through acting on any principles they choose, there is no way to justify the good of the sense of justice. What Rawls needs to establish is that the exercise of the sense of justice must embody a shared final end, namely the realization of our nature as free being, which is a higher-order good that every rational agent has reason to pursue. But if the realization of human nature becomes a common end, does it not imply that justice as fairness is a teleological theory? This claim sounds implausible because it is well known that the aim of *A Theory of Justice* is to develop a deontological conception of justice as fairness to replace utilitarianism which is the most prominent representative of teleological theory. However, I believe that this claim not only makes sense, but also properly accounts for Rawls's later philosophical turn to political liberalism. In the following section, I will defend this claim.

## 7 Justice as Fairness as a Teleological Theory

A teleological theory consists of two components: "the good is defined

independently from the right, and then the right is defined as that which maximizes the good." (TJ:24/21-22 rev.) It is important to note that the good is referred to non-moral, independently identifiable value. This is because, according to Frankena, to allow "the moral quality or value of something to depend on the moral value of whatever it promotes would be circular."[63] Teleological theories then make the right dependent on the non-morally good. In order to know whether something (an action, a policy, or an institution) is just, one must first know what is good in the non-moral sense, and whether the thing in question can maximize, or is intended to maximize the good. These non-moral goods can be variously identified with happiness, pleasure, human excellence, knowledge etc. We can then have different teleological theories such as ethical egoism, perfectionism, hedonism and utilitarianism. What they commonly share is the idea of the maximization of the good. Rawls further thinks that this is the deepest appeal of teleological theories because it seems to embody the idea of rationality. "It is natural to think that rationality is maximizing something and that in morals it must be maximizing the good." (TJ:24-25/22 rev.) By definition, deontological theory is the very opposite of teleology in that it neither specifies the good independently of the right, nor interprets the right as maximizing the good. (TJ:30/26-27 rev.) Rawls believes that justice as fairness meets the requirement of deontology in these two respects.

Rawls agrees with utilitarianism about the definition of the good as the

---

[63] Williams Frankena, *Ethics* (Englewood Cliffs, New Jersey: Prentice Hall, 1973), second edition, p.14. It is J. H. Muirhead who first uses the teleological/deontological division to describe all ethical theories; see his *Rule and End in Morals* (Oxford: Oxford University Press, 1932). Rawls does not use the term "non-moral good," but he stresses that the goodness of things is a separate class which is judged without referring to what is right. (TJ:25/22 rev.)

satisfaction of rational desire, or the satisfaction of rational plans of life. (TJ:408/358-359 rev.) What he objects to is the utilitarian claim that the satisfaction of any desire has some value in itself which must be taken into account in deciding what is right. This is unreasonable because not all desires are legitimate. Racial discrimination in a mainly white society, for instance, may bring a lot of pleasure to the white. However, this kind of pleasure should not be given equal weight to other preferences in the utility calculations because it violates minorities' rights. No matter how much satisfaction is derived from those desires, they have no moral weight whatsoever. A reasonable political morality, Rawls argues, should incorporate the concept of right as prior to that of the good. It follows that the principles of justice put limits on which satisfactions have value. They should impose restrictions on what are reasonable conceptions of one's good from the outset. Hence, "in justice as fairness one does not take men's propensities and inclinations as given, whatever they are, and then seek the best way to fulfil them." (TJ:31/27 rev.) This is the first reason why justice as fairness is a non-teleological theory. The second reason is that the idea of maximization does not play any role in justice as fairness. Although the parties in the original position want to secure as many social primary goods as possible, they would choose a principle of equal liberty and restrict social and economic inequalities to the greatest benefits of the least advantaged. So, "there is no reason to think that just institutions will maximize the good." (TJ:30/27 rev.) Since both essential features of teleological theory do not apply to justice as fairness, it is therefore identified as a deontological theory.

Rawls's characterisation of teleology and deontology has been widely accepted. His criticism of utilitarianism implies that all kinds of teleological

theories are doomed to failure. However, I believe this is mistaken. Rawls's Kantian argument for congruence is actually a type of teleological theory which need not face the difficulties of utilitarianism. I want to make two points to support this claim. First, I will show that Rawls's argument is grounded on a teleological account of human nature which makes the concept of the good prior to that of the right. Second, I will argue that the idea of maximization is not a defining feature of teleological theories. In other words, a conception of justice grounded on a certain ideal of human nature without adopting the maximizing principle is still a teleological theory. Justice as fairness is such an example.

To validate my first point, Rawls's argument is stated as follows:

> 1. The good is defined by the satisfaction of rational desire.
>
> 2. It is rational for us to have a higher-order desire to express our nature as free and equal rational beings. (TJ:574/503 rev.)
>
> 3. To express our nature as a being of a particular kind is to act on the principles that would be chosen if this nature were the decisive determining element. (TJ:253/222 rev.)
>
> 4. Rawls's two principles of justice would be chosen by the parties in the original position which characterizes our nature as free and equal rational beings.
>
> 5. Therefore, acting in accordance with the principles of justice is something that we have a regulative desire to do. The desire to act justly and the desire to express our nature turn out to be the same desire. So, acting from our sense of justice is a regulative good for rational persons. (TJ:572/501 rev.)

This argument noticeably depicts a teleological outlook. What is right is

defined by the full expression of our nature. Our nature is our *telos*. It can only be fully realized and perfected through the effective exercise of the sense of justice. This should not surprise us because we have already seen that Rawls is a perfectionist according to the Kantian interpretation, and perfectionism is a type of teleology. [64] There are some distinct features of Rawls's teleology which allow Rawls to avoid those difficulties faced by utilitarianism.

First, Rawls's theory can avoid the problem of illegitimate preferences. It does not hold that the satisfaction of desires stems from the same source, and therefore ranks equally without any independent discrimination. Rather, it takes the desire to realize our nature as a qualitatively different desire. As Rawls puts it, "in order to realize our nature we have no alternative but to plan to preserve our sense of justice as governing our other aims." (TJ:574/503 rev.) It is a higher-order regulative desire. When other desires are in conflict with the sense of justice, the latter has absolute priority. Since the sense of justice is characterized by Rawls's principles of justice, so any preferences that violate the principle of equal liberty will be disallowed at the very beginning.

Moreover, the calculation of the greatest sum of satisfaction does not have any place in deciding what is right or wrong in Rawls's case. The principles of justice are justified by its realization of freedom and equality. The possibility that the loss of freedom for some is justified by a greater good shared by others will not arise because a right to equal liberties is grounded on the Kantian conception

---

[64] Rawls holds that there are two variants of perfectionism. The first one is that it is the sole principle of a teleological theory; the second one is that it is one of several principles in an intuitionist theory. The principle of perfection is balanced against others by intuition. (TJ:325/285-286 rev.)

of the person. This conception takes the separateness of individuals seriously. However, this is not a deontological view as Sandel describes: "it describes a form of justification in which first principles are derived in a way that does not presuppose any final human purposes or ends, nor any determinate conception of the human good."[65] For according to the Kantian interpretation, justice as fairness is justified by appealing to a particular vision of human *telos*. Rawls repeatedly reminds us that his whole theory is grounded on a conviction that we have a higher-order interest to realize our free nature. His principles of justice are constructed to represent and realize this fundamental interest.[66] The primacy of justice is founded on a final human end. Once the overriding interest in realising our nature is granted, a teleological theory warrants the priority of justice.[67]

Sandel argues that this kind of teleological liberalism will put the primacy of justice at risk. For "where the right is instrumental to the advancement of some end held to be prior, the denial of liberty for some may be justified in the name of an overriding good for others."[68] This concern brings out the second contrast between Rawls's teleology and utilitarianism. As Frankena describes it, utilitarianism regards the right simply as a means to promote an independent and non-moral good. Virtuous or right actions are merely instrumental to the

---

[65] Sandel, *Liberalism and the Limits of Justice*, p.3.

[66] For instance, in the revised edition of *Theory of Justice*, Rawls remarks that "the basic rights and liberties and their priority are there said to guarantee equally for all citizens the social conditions essential for the adequate development and the full and informed exercise of their two moral powers—their capacity for a sense of justice and their capacity for a conception of the good—in what I call the two fundamental cases." (TJ:xii, rev.) This idea has been fully elaborated in his "Kantian Constructivism in Moral Theory," *Collected Papers*, pp.303-58.

[67] Sandel actually admits that it is possible to establish the moral priority of justice without recourse to deontology. Mill's liberalism is a case in point. Sandel, *Liberalism and the Limits of Justice*, p.3.

[68] Sandel, *Liberalism and the Limits of Justice*, p.18.

maximization of the good. When the claim of right conflicts with the aggregation of non-moral good, the former ought to give way to the latter. Rawls rejects this view by stating that acting justly is constitutive to the perfection of human nature. The desire to be a just person and the desire to express one's nature move a person to act in the same way. Doing what justice requires is not something separated from our good. On the contrary, it is an intrinsic value essential to our well-being. So, the problem of sacrificing the right for the good will not happen in Rawls's case because doing so is irrational.

Someone may ask if there is such a great contrast between Rawls's proposition and utilitarianism, whether it is proper to label both of them as teleological theories? Korsgaard suggests that they are teleological in a different sense. Rawls's position is closer to classical teleological ethics represented by Aristotle while utilitarianism is a type of modern teleological theory. Although both share the same view that the good is realized through virtuous action, "classical teleologists argue that virtue is identical with the best state of a human being, while modern ones argue that virtue promotes an independent, nonethical good."[69] Rawls belongs to the classical camp because first, acting justly is constitutive to the perfection of human nature, and second, the realization of our nature is experienced as a moral good. Korsgaard is unsatisfied with Frankena and Rawls's differentiation of ethical theories into teleological and denotological, and assigning classical Greek theories along with utilitarianism to the first category and Kant's theory to the second. For this widely accepted distinction has not only

---

[69] Christine Korsgaard, "Teleological Ethics," in E. Craig ed., *Routledge Encyclopedia of Philosophy* (London: Routledge, 1998), retrieved from http://www.rep.routledge.com/article/L103. Although Korsgaard does not mention Rawls in this essay, I believe that her account of the classical view provides a reasonable interpretation of Rawls's congruence argument.

obscured the fundamental difference between classical and modern teleological theories, but also neglected the similarity between Kant and classical theory.[70] Following Korsgaard's categorisation, there is no doubt that Rawls is a classical teleologist.

It may also be argued that even if justice as fairness gives priority to the concept of the good, it is still deontological because it denies the idea of maximization. Nevertheless, it is wrong to view maximization as a defining feature of teleology. Extending the principle of rational choice for one man to society as a whole, and pursuing the greatest net balance of satisfaction summed over all the individuals is a feature of utilitarianism, but not of teleology. Other teleological theories need not adopt this requirement. Take Aristotle as an example. For Aristotle, happiness (*eudaimonia*) is what rational agents ultimately desire. To lead a happy life depends on the realization of our distinctive human function as rational being through the practice of virtues. But the idea of maximization does not play any role in his thought. Although a political community must provide a congenial environment to enable citizens to lead a good life, it has no duty to maximize different kinds of virtue among citizens. Nor does Aristotle hold that we should compromise someone's good for the sake of the greater happiness of others. There is no such maximizing formula in judging when an action is virtuous or not.

There are at least three reasons to explain why a teleologist need not accept the concept of maximization. The first is concerned with the source of value.

---

[70] A major collection of essays offering a serious challenge to the traditional distinction about ancient and modern ethics can be found in S. Engstrom and J. Whiting (eds.) *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty*, (New York: Cambridge University Press, 1996).

Maximization presupposes that all values can be measured in accordance with a common scale. Their qualitative difference is thus entirely neglected because they are supposed to be reducible to the ultimate good. This is value monism that utilitarianism presumes. However, this meta-ethical view about value is questionable. For a teleological pluralist who believes that values are ultimately incommensurable, the very idea of maximization does not make any sense in the first place. He could argue that while there is a diversity of intrinsically valuable goods, they can only be realized in their distinct spheres according to their internal logic.[71] When there is conflict between different types of reason, no conclusive reason is available that one reason is necessarily outweighed by another because of the lack of a common commensurable criterion.

The second reason is related to the nature of value. According to Rawls's understanding of teleology, once a good has been independently defined, what we should do is to promote it as far as possible. Rationality dictates that we ought to prefer the better to the lesser, and ought to do what is best. What is best is to maximize the good. But there are some goods to which the idea of maximization may not appropriately apply. Friendship, for instance, is unquestionably an important good for our life. Should we have a duty to maximize friendship then? Provided that we should, this statement can be understood in two different ways. It could mean either that we should make every effort to make as many new friends as possible, or that we should perfect our relationship with existing friends. These two ways to promote friendship are not conceptually incompatible with

---

[71] For different accounts of value pluralism, see Isaiah Berlin, *Four Essays on Liberty* (Oxford: Oxford University Press, 1969), p.171; Thomas Nagel, *Mortal Questions*, (Cambridge: Cambridge University Press, 1979), pp.128-41; Michael Stocker, *Plural and Conflicting Values* (Oxford: Clarendon Press, 1990).

each other. We can do both at the same time. But they connote very different senses of "more is better." The first way only takes numbers into account. If we care about the *quality* of friendship, what we should do is perhaps to devote more time and love to our friends, and pay more attention to their needs. From a utilitarian point of view, there is nothing wrong with betraying one's friends for a much larger number of new friends if this can boost the net balance of one's happiness. Friendship, however, cannot be exchanged this way. It is absurd to demand that a person should give up a special relation with his friend at any time when the sums come in from the utility network. What is at stake is not the accuracy of the calculation, but the calculation's distortion of the very nature of friendship. True friendship involves loyalty, trust and commitment which cannot be overridden by utilitarian calculation.[72] In this case, a teleologist need not accept maximization as a necessary means to promote the goodness of friendship.[73]

Finally, Rawls's account of human nature has a built-in element to counter the trend of maximization. We know that Rawls's major complaint about utilitarianism is that it does not take seriously the plurality and separateness of individuals because it conflates the desires of all persons into a coherent system. As a consequence, the loss of freedom for some is morally justified by a greater good shared by others. However, since Rawls views freedom and equality as an essential property of our nature, we have a fundamental interest in protecting our

---

[72] This is similar to Williams's critique of utilitarianism as an attack on a person's integrity. See "A Critique of Utilitarianism" in *Utilitarianism: For and Against* ed. J. J. C. Smart & B. Williams (Cambridge: Cambridge University Press, 1973), pp.116-7.

73 I am indebted to Tang Siu-fu and Tang Wai-sang for this point.

equal liberties in deciding the principles of social cooperation for reciprocity.[74] Maximization is unacceptable because it will impair our identity as free and equal beings. In Rawls's case, therefore, it is exactly because we have a regulative desire to realize our nature that we will not adopt the concept of maximization.

To conclude, my above argument shows that maximization of the good is not a defining feature of teleology. What is essential to a teleological political theory is that it justifies principles of justice with reference to some final purpose or the realization of human nature.[75] This is the main reason I ascribe Rawls's argument for congruence to a view of classical teleology. If my argument is valid, the widely accepted dichotomy of teleology and deontology then collapses. Here we come to the last concern of this chapter: can this teleological theory offer a successful argument for congruence which is consistent with desire-based deliberative rationality?

## 8 The Difficulty of Congruence

In this section, I will show that there is a deep tension between teleology and the desired based rationality. This tension results in the failure of congruence and urges Rawls to recast his theory as a political conception of justice.

The primary aim of congruence is to vindicate the good of the sense of justice from the point of view of prudential rationality so that the priority of moral motivation can be secured. A conception of justice is sufficiently stable if "from

---

[74] It is, however, quite strange for Rawls to say that equality is an essential property of human beings. It should be a comparative concept concerning the relations between persons.

[75] Korsgaard, "Teleological Ethics."

the standpoint of the individual, the desire to affirm the public conception of justice as regulative of one's plan of life accords with the principles of rational choice." (TJ:577/505 rev.) More importantly, when rational individuals judge the desirability of the sense of justice, they "assess their situation independently from the constraints of justice."(TJ:399/350 rev.) Putting congruence in the context of the Kantian interpretation, what Rawls strives to establish is that it is always rational for members in a well-ordered society, who have full knowledge of their conceptions of the good, to regard acting justly as a supremely regulative good because doing so can effectively express their nature as free and equal rational being. Stability is grounded on a combination of rationality and teleology.

The crucial question is whether deliberative rationality can have such normative force as to lead individuals to accept and act upon Rawls's account of human nature. We know that a plan of life is rational when it satisfies what we most want after informed deliberation. Our good is determined by the plan of life that we would adopt with full deliberative rationality. Thus an individual would be acting rationally if he would be doing what, all things considered, he wants to do most. This is a desire-based means-end conception of rationality. Whether an action is rational depends on whether it is the best means to satisfy our given fundamental desires. However, rationality itself cannot dictate which end we should have in the maximal class of ends. The end is the result of our choice. Given that persons are left free to choose, it is inevitable that "individuals find their good in different ways, and many things may be good for one person that would not be good for another." (TJ:448/393 rev.) It is a permanent feature of modern society that rationality is coexistent with a plurality of conceptions of the good.

This account of rationality poses a serious challenge to Rawls: if people do not care about justice after careful and informed deliberation, there is no ground for him to further argue that it is rational for them to do so. The internal logic of desire-based rationality restrains Rawls from asserting the universal acceptance of the sense of justice as a higher-order end that rational agents would share. Rawls is on the horns of a dilemma. On the one hand, rationality cannot determine what common final ends all of us ought to have. People's ends are inevitably diverse in a free society. On the other hand, to support his argument for stability, Rawls needs to show that people will commonly endorse the sense of justice as a regulative good embedded in their plans of life. A deep tension exists between pluralism and the priority of the sense of justice. In order to justify congruence, Rawls has taken the second horn by arguing that regardless of their different aims and desires, rational people have a shared higher-order end to realize their nature as free and equal rational being. This desire has absolute priority over other desires. As Rawls explains,

> Therefore, in order to realize our nature we have no alternative but to plan to preserve our sense of justice as governing our other aims. This sentiment cannot be fulfilled if it is compromised and balanced against other ends as but one desire among the rest. (TJ:574/ 503 rev.)

When Rawls appeals to this teleological account of human nature, he of course believes that he is making an objectively true statement. Thus to act upon his principles is equivalent to realising our most fundamental good. It is rational for people to adopt a Kantian conception of the good to lead their life. Rawls later patently acknowledges this assumption:

> This is the premise that in the well-ordered society of justice as fairness, citizens hold the same comprehensive doctrine, and this includes aspects of Kant's comprehensive liberalism, to which the principle of justice as fairness belong. (PL:xlii)

Having acknowledged this assumption, however, it is inconsistent with deliberative rationality and the fact of pluralism. After all, "it is rational for members of a well-ordered society to want their plans to be different." (TJ:448/393 rev.) There is no way to expect that rational people would commonly accept the same Kantian worldview in a well-ordered society. Rationality parts company with teleology. The real problem is not that Rawls's account of human nature is unconvincing, or that Rawls overlooks the fact of pluralism. It is rather that desired-based rationality is unable to support teleology in a pluralistic society. This internal inconsistency has left little room for congruence. Surprisingly, after going to great lengths, Rawls concedes the bounds of rationality by saying that:

> To justify a conception of justice we do not have to contend that everyone, whatever his capacities and desires, has a sufficient reason (as defined by the thin theory) to preserve his sense of justice. For our good depends upon the sort of persons we are, the kinds of wants and aspirations we have and are capable of. It can even happen that there are many who do not find a sense of justice for their good. (TJ:576/504-05 rev.)

This is equivalent to admitting that for those who do not accept the Kantian conception of person, we cannot recommend the sense of justice as a good to them since they do not have sufficient reason to do what justice requires. They may still comply with justice for prudential reasons. But they will not deem the sense of justice a regulative good of their plans of life. In other words, without endorsing the Kantian interpretation of human nature, there is no other reason to

justify the goodness of justice. The possibility of congruence is subject to a rather bold assumption that the majority of people would accept a Kantian conception of the good. In a liberal society where people have different kinds of wants and plans, Rawls's hope for congruence is fundamentally unrealistic.

What then might Rawls say to those who do not find it a good for them to act justly? Of course Rawls could not dismiss them as irrational. There is no room for Rawls to make such a claim. Against all expectations, Rawls's answer is that

> It is, of course, true that in their case just arrangements do not fully answer to their nature, and therefore, other things equal, they will be less happy than they would be if they could affirm their sense of justice. But here one can only say that their nature is their misfortune...Under such conditions penal devices will play a much larger role in the social system. (TJ:576/504 rev.)

This charge against those rational non-Kantian people is inconsistent with Rawls's conviction of justice. If their failure in recognising the good of the sense of justice grows out of their unfortunate nature, why should they be responsible for that? According to Rawls, one's natural endowment is the outcome of the natural lottery, and is arbitrary from a moral point of view. We know that this is the main moral ground for his difference principle. Following this logic, those suffering from unhappiness owing to a nature not of their choosing should not be penalized. On the contrary, they are even entitled to some kind of compensation.[76]

Furthermore, as I have thoroughly demonstrated in the first chapter of this

---

[76] Matt Matravers has raised this issue in *Justice and Punishment* (Oxford: Oxford University Press, 2000), pp.145-47.

thesis, Rawls's retreat to penal devices to ensure strict compliance has confused the distinction between moral stability and social stability. What Rawls aims to achieve is the motivational priority of the sense of justice. If this project fails and a certain portion of members lack sufficient motive to act in accordance with justice as fairness, the employment of coercive force is necessary to secure social stability according to Rawls. However, in this case the use of state power for social stability actually indicates the failure of moral stability. The pursuit of congruence is not a means to social stability, but an independent consideration for the desirability of justice as fairness. When Rawls concedes that there are no sufficient reasons to convince every rational individual to see the good of the sense of justice, it is not a practical matter, but a matter concerning the justifiability of justice as fairness. The later Rawls is fully aware of this distinction:

> To clarify the idea of stability, let us distinguish two ways in which a political conception may be concerned with it. In one way we view stability as a purely practical matter...As long as the means of persuasion or enforcement can be found, the conception is viewed as stable. But, as a liberal conception, justice as fairness is concerned with stability in a different way. Finding a stable conception is not simply a matter of avoiding futility. Rather, what counts is the kind of stability, the nature of the forces that secure it. (PL:142)

Only against this background can we understand why the later Rawls stresses that the problem of stability is fundamental to political philosophy and why an inconsistency internal to Part III of *Theory of Justice* has urged him to make basic readjustments of his whole enterprise. "All differences are consequences of removing that inconsistency." (PL:xviii) For example, Rawls has totally dropped

the Kantian interpretation and the idea of congruence in *Political Liberalism*. While he retains the theory of deliberative rationality, he has introduced the idea of reasonableness to constrain the use of rationality. The idea of an overlapping consensus is proposed as a replacement for congruence in justifying stability for a democratic society characterized by reasonable pluralism. Therefore, to evaluate Rawls's later philosophical development, the most important frame of reference is to consider whether his political liberalism can provide a better justification for moral stability. This is what I am going to do in the next chapter.

## CHAPTER 5

# THE LIMITS OF AN OVERLAPPING CONSENSUS

In the previous chapter, I have argued that a tension between rationality and the Kantian interpretation of justice as fairness has failed the congruence argument for stability. In order to resolve this tension, Rawls makes a paradigmatic shift to political liberalism, and rests stability on the idea of an overlapping consensus. Rawls believes that a freestanding liberal political conception of justice will be endorsed by a plurality of reasonable comprehensive doctrines from different perspectives. The priority of political values is secured by the consensus. This chapter sets out to examine the plausibility of this approach. I shall raise my doubts about whether a political conception of justice can win sufficient support of comprehensive doctrines.

It should be noted that this chapter does not aim to offer a comprehensive evaluation of political liberalism. My discussion will only focus on a specific question: to what extent will a political conception of justice provide sufficient reason for a rational agent, who may hold a non-liberal conception of the good, to accept the priority of justice? This is the central question of political liberalism. Its answer will determine the degree of stability of a conception of justice. What I provide is basically an internal critique. I will not challenge the desirability of justice as fairness as a political conception from outside. My major concern is whether Rawls's dividing our values and motives into two distinct parts is a wise way to solve the problem of moral stability. My argument will show that it is not a right direction. This will then pave the way for my idea of potential congruence, the main theme of the last chapter.

This chapter is divided into four sections. Section 1 introduces the basic ideas of political liberalism and shows how the problem of stability is tackled by the idea of an overlapping consensus. Section 2 examines Rawls's first argument for the possibility of political liberalism which is concerned with the greatness of political values. I contend that this argument is insufficient to establish the priority of political values over non-political ones. Section 3 takes up the second argument that justice as fairness as a political conception will be the focus of an overlapping consensus. I shall use Kantianism and utilitarianism as examples to demonstrate the difficulties of this approach in justifying the overridingness of political values. Finally, section 4 focuses on the idea of burdens of judgment and argues that it presupposes a moderate scepticism which will substantially undermine political liberalism.

## 1 The Idea of a Political Conception of Justice

In *Political Liberalism*, Rawls states that the idea of an overlapping consensus is primarily designed to replace the Kantian interpretation of justice as fairness to resolve the problem of stability in face of the challenge of reasonable pluralism in modern society. It is for this reason that he transforms justice as fairness into a freestanding political conception. (PL:xlii-xliii) It is, therefore, a good idea to start with an analysis of the internal relation between an overlapping consensus and stability.

According to Rawls, a conception of justice would be signalled as the focus of an overlapping consensus when it is endorsed by citizens who affirm fundamentally different and opposing, though reasonable, comprehensive

doctrines. The term "overlapping" denotes convergence on a conception of justice from different perspectives for different reasons. Rawls thus says:

> In such a consensus, the reasonable doctrines endorse the political conception, *each from its own point of view.* Social unity is based on a consensus on the political conception; and stability is possible when the doctrines making up the consensus are affirmed by society's politically active citizens and the requirements of justice are not too much in conflict with citizens' essential interests as formed and encouraged by their social arrangement. (PL:134, emphasis added)

Some salient features arising from this idea are noteworthy. First of all, in reaching a consensus, citizens need not abandon their reasonable conceptions of the good. They decide for themselves how the political conception is related to their comprehensive worldviews. It is their autonomy to "view the political conception as derived from, or congruent with, or at least not in conflict with, their other values." (PL:11) The motivating reason for obedience varies from person to *person* depending on their philosophical and religious beliefs. In this sense, political liberalism is more tolerant and flexible than other theories of justice.

That being said, political liberalism demands that citizens give precedence to political values over non-political values in case of conflict. This requirement of priority is a key issue of political liberalism. For Rawls's theory presumes a dualism between the point of view of the political conception and the many points of view of comprehensive doctrines. As Rawls puts it, "citizens' overall views have two parts: one part can be seen to be, or to coincide with, the publicly recognized political conception of justice; the other part is a (fully or partially)

comprehensive doctrine to which the political conception is in some manner related." (PL:38) Both political and non-political parts are important in citizens' lives. On the one hand, as citizens, they participate in political activities and affirm the values of political justice. They will also use political values to judge political institutions and social policies. On the other hand, they have their non-political aims and commitments. These commitments give shape to a person's way of life. They may constitute what Williams called a person's ground project which provides "the motive force which propels him into the future, and gives him a reason for living."[1] Rawls fully acknowledges the existence of these two independent standpoints. Our political identity defined by our moral powers as free persons are distinct from our non-political identity defined by our ends and projects.

Once our life is divided into two parts, the possibility of radical conflict is also there. When this happens, how the political values can overrule non-political ones becomes a salient problem. This is a difficult but extremely important issue that political liberalism must face. Political liberalism must give reasons to convince citizens that they should honour the demand of justice even if doing so may sacrifice their fundamental interests. The priority must be given to political values even from a citizen's comprehensive point of view. Rawls sets a daunting task for himself. On the one hand, he takes reasonable pluralism seriously and accepts a desire-based view of practical reason that motivational force for actions must stem from one's comprehensive doctrines; on the other hand, he aims to

---

[1] Bernard Williams, "Persons, Character and Morality" in *Moral Luck* (Cambridge: Cambridge University Press, 1981), p.13.

justify the priority of political values. Rawls believes that the problem will be solved by constructing a political conception of justice that can be the focus of an overlapping consensus among reasonable conceptions of the good.

A political conception of justice has three distinctive features. First, it is a moral conception worked out for the basic structure of modern constitutional democracy. The basic structure refers to a society's main political, social and economic institutions. (PL:11) Second, a political conception is *presented* as a freestanding view, meaning that "it is neither presented as, nor as derived from, such a [comprehensive] doctrine applied to the basic structure of society, as if this structure were simply another subject to which that doctrine applied." (PL:12) It involves no wider commitment to any general and comprehensive moral ideals.[2] Rather, it presents itself as a module that fits into and can be supported by various reasonable comprehensive doctrines. Third, the content of a political conception is "expressed in terms of certain fundamental ideas seen as implicit in the public political culture of a democratic society." (PL:13) These ideas are supposed to be widely shared and independent of any particular comprehensive doctrine. Among them, the most fundamental idea is that of society as a fair system of cooperation among free and equal citizens. Rawls hopes that a political conception of justice may be developed out of these shared ideas, and therefore gain the support of an overlapping consensus.

The central ideas of political liberalism can be summarized as follows. A conception of justice is stable when the motivational priority of the sense of

---

[2] A moral conception is general if it applies to a wide range of subjects. It is comprehensive if it includes a wide range of values and virtues in human life. (PL:13)

justice is affirmed. Given the permanent fact of reasonable pluralism of modern societies, a just and stable order is possible only if the basic structure is effectively regulated by a conception of justice that can claim to be the focus of an overlapping consensus. To realize this goal, the conception of justice must present itself as freestanding without depending on any comprehensive doctrines. Rawls contends that justice as fairness should be presented as a political conception in terms of its scope and the source of justification. Two puzzling questions immediately follow. First, how can a political conception *present* itself as freestanding while being understood as part of, or derivable within, a comprehensive doctrine? Moreover, what kind of justificatory force would an overlapping consensus add to a freestanding political conception that has already been justified by appealing to values implicit in the public culture?[3] In reply to these questions, Rawls clarifies that justification of justice as fairness should be understood as consisting of two stages:

> In the first stage it is worked out as a freestanding (but of course moral) conception for the basic structure of society. Only with this done and its content—its principles of justice and ideals—provisionally on hand do we take up, in the second stage, the problem whether justice as fairness is sufficiently stable. Unless it is so, it is not a satisfactory political conception of justice and it must be in some way revised. (PL:141)

There is an internal division of labour between the two stages. In the first stage citizens are ignorant of their conceptions of the good, and can only consider

---

[3] For example, Samuel Scheffler has pointed out the ambiguity of Rawls's account of the first question while the second question has been raised by Jurgen Habermas. See Scheffler, *Boundaries and Allegiances* (New York: Oxford University Press, 2001), pp.138-39; Habermas, "Reconciliation through the Public Use of Reasons: Remarks on John Rawls's Political Liberalism," *Journal of Philosophy*, 92 (1995), pp. 119-22.

those political values implicit in the public culture. These values are then modelled into the original position from which Rawls's principles of justice are derived. Since citizens make their decision behind the veil of ignorance and will only take political values into account, the content of justice will not favour any comprehensive doctrines. It is freestanding in this sense. Rawls calls this stage *pro tanto* justification. But the process of justification is unfinished because the political values are only part of citizens' overall view. While the political standpoint plays an essential role in elaborating a political conception of justice for fair cooperation, it is not by default overriding relative to one's comprehensive doctrine. It may be overridden by citizens' comprehensive doctrines once all values are counted. (PL:386) Its precedence depends on how it is related to a citizen's comprehensive value system.

This concern over priority points to the second stage justification in which "it is left to each citizen, individually or in association with others, to say how the claims of political justice are to be ordered, or weighed, against nonpolitical values." (PL: 386) The idea of an overlapping consensus comes into play in this stage. Since citizens are allowed to have full knowledge of their worldviews, their practical reasoning will take both political and non-political values into account. Citizens individually decide for themselves in what way the political conception is related to their more comprehensive views. Rawls thus concludes that "even though a political conception of justice is freestanding, that does not mean it cannot be embedded in various ways—or mapped, or inserted as a module—into the different doctrines citizens affirm." (PL:387)

Rawls further distinguishes two types of justification in the second stage: *full*

*justification* by individuals and *public justification* by political society. The former refers to the endorsement of a political conception from an individual's comprehensive perspective. A political conception is fully justified when an individual accepts it by relating its principles in some way to his comprehensive doctrine as either true or reasonable. Since citizens affirm different reasonable conceptions of the good, the justificatory reasons will correspondingly vary from person to person. It is possible that someone considers the political conception completely justified while others find it entirely ungrounded. (PL:386) Nevertheless, like *pro tanto justification*, full justification does not yield stability for the right reason in a pluralistic society because there is no agreement among citizens on the authoritative status of a political conception. "It is left to each citizen, individually or in association with others, to say how the claims of political justice are to be ordered, or weighed, against nonpolitical values." (PL:386)

What political liberalism strives for is public justification which occurs "when all the reasonable members of political society carry out a justification of the shared political conception by embedding it in their several reasonable comprehensive views." (PL:387) In contrast with full justification, it depicts a situation in which different reasonable comprehensive doctrines converge on a freestanding political conception. It is not a result of political bargaining. Rather, it is subject to citizens who *individually* decide how the political conception is related to their comprehensive views. "In some cases the political conception is simply the consequence of, or continuous with, a citizen's comprehensive doctrine; in others it may be related as an acceptable approximation given the circumstances of the social world." (PL:xxi) We may say public justification is not

determined by Rawls or any particular individuals. Rather, it is a result of the union of different kinds of full justification. The mutual recognition of the existence of such a consensus then offers sufficient motivation for each individual to comply with the political principles. Therefore, as far as the justificatory status of an overlapping consensus is concerned, Rawls argues that it is a necessary condition of public justification:

> There is, then, no public justification for political society without a reasonable overlapping consensus, and such a justification also connects with the ideas of stability for the right reasons as well as of legitimacy. (PL:388-89)

This means that justice as fairness is not publicly justified and therefore legitimate until the free-standing conception is shown to be sufficiently stable by the fact of an overlapping consensus. The arguments in both stages are essential to the justifiability of Rawls's principles of justice.[4] Without the first stage citizens are unable to work out a freestanding conception of justice as a basis for fair social cooperation. Its substantive content is entirely given by the political argument. Without the second stage citizens do not know whether the political values embodied in the political conception can occupy a proper and overriding place in their comprehensive ethical outlook. The motivational priority of justice is ultimately confirmed by the argument of the second stage. Furthermore, we should note that the arguments of both stages are available to citizens. It is citizens themselves who take up political and non-political standpoints and decide their proper relations.

---

[4] This issue has been thoroughly discussed in Chapter 3.

We now see that having recognized the permanent fact of reasonable pluralism as a result of the free exercise of human reason, the later Rawls has given up the ambition of grounding stability on congruence by resorting to a Kantian interpretation of human nature. He believes that this approach is empirically impractical and morally illegitimate. A comprehensive conception of the good, including Kantian liberalism, can be maintained only by the oppressive use of state power. This contradicts the liberal principle of legitimacy according to which "our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason." (PL:137) Rawls concedes the inability of rationality to vindicate the good of the sense of justice in a pluralistic society.

However, this does not mean that the problem of stability has been resolved. On the contrary, the question of motivational priority becomes more salient. How can citizens commonly develop an effective sense of justice to honour political principles if they believe in a plurality of incompatible conceptions of the good? Rawls's strategy is to draw a sharp distinction between values in the political domain and those in the non-political one. These two domains need not be in union. As Rawls puts it, "it can happen that in their personal affairs, or in the internal life of associations, citizens may regard their final ends and attachments very differently from the way the political conception supposes." (PL:31) The split of the self into two parts unavoidably results in tension. The conflict can happen within oneself. A person's ground projects may be in deep conflict with the requirement of justice. It can also take place between two persons. People lead

their lives in accordance with their worldviews. If they have fundamentally different conceptions of the good, they will naturally have different views on how the social world should be arranged. Even if they uphold the same set of political principles, it could be out of radically different reasons. In a nutshell, the most serious challenge to political liberalism is how it can justify the priority of political values over non-political values. Therefore, the possibility of political liberalism hinges on answering the following question: "how can the values of the special domain of the political—the values of a subdomain of the realm of all values—normally outweigh whatever values may conflict with them?" (PL:139) In the rest of this chapter, I will focus on this issue and argue that political liberalism fails to offer a satisfactory argument for the priority of political values.

## 2 The Importance of Political Values

Rawls's argument consists of two complementary parts. The first part states that political values are very great values and hence not easily overridden. (PL:139) The second part says that there are many reasonable ways in which the wider realm of non-political values is positively related to the values of the political domain so that an overlapping consensus is possible. (PL:140) Rawls believes that when both conditions are met, a conception of justice will be stable. I shall argue that both parts are problematic. This section will examine the argument of the first part.

The first argument seems simple and straightforward. Rawls believes that since the political values expressed by the conception of justice are very important, they thus have sufficient weight to override all other values that come into conflict

with them. These values include political and civil liberties, fair equality of opportunity, economic reciprocity, and the social bases of mutual respect between citizens. They also cover the values of public reason expressed in the guidelines for public inquiry and the precepts governing reasonable political discussion.

One possible way to justify the importance of political values is to appeal to its object of application. They apply to the basic structure of society and specify the fundamental terms of social cooperation. The basic structure has a profound impact on every citizen's life prospect from the start. It defines our fundamental rights and duties, and determines our initial place in life; it also shapes our plans of life and limits our ambitions in different ways. Citizens recognize the special status of the basic structure and therefore assign top priority to political principles.

Although this account provides a *general* justification for the precedence of political values applied to the basic structure, it does not establish the overridingness of Rawls's political conception of justice. Other competing conceptions of justice could agree on the supreme importance of the basic structure while denying Rawls's specific account of political values as overriding. Libertarianism, for example, may contend that the right to self-ownership and the right to private property should be the most important values for the basic structure. Rawls needs a more substantive argument to vindicate his claim. He has indeed done so. For example, he reminds us that his account of fair cooperation among free and equal person, arbitrariness of natural endowment and social circumstance from the moral point of view, and people's higher-order interest in developing their two moral powers adds up to a liberal egalitarian ideal. However, in *Political Liberalism*, Rawls reminds us that these ideas do not stem from any

comprehensive liberal worldview. Rather, they are seen as implicit in the public political culture of a democratic society. The political conception of justice is constructed out of these shared fundamental ideas without appealing to any comprehensive moral doctrines. This seems to imply that the importance of political values largely depends on a sociological fact that citizens do take these political ideas very seriously. Their normative force is explained by their prominence in a particular form of political culture.

Rawls wants to avoid any controversy over conceptions of the good so that the conception of justice is given a chance to be the focus of an overlapping consensus. This strategy, though understandable, makes for a weak argument. There is nothing wrong for a political theory to rest its justification on values in the public culture. However, in a pluralistic society we may reasonably doubt how plausible it is that citizens can have a high level of consensus on a set of political ideas and at the same time maintain deep disagreement on their conceptions of the good. If a plurality of reasonable yet incompatible comprehensive doctrines is a normal result of the exercise of human reason, why does the same situation not apply to the political sphere? Our public culture contains as many competing political ideas as comprehensive doctrines. Most political values are essentially contested concepts. For instance, people have reasonable disagreement about the proper meaning of freedom and equality, and their priority in a political system. Nor is Rawls's idea of society as a fair system of cooperation uncontroversial. As a matter of fact, we have witnessed the predominance of the New Right in American society in the past two decades. The rightists obviously do not share Rawls's ideas of social cooperation and moral personality. Rawls may argue that the New Right's interpretation of political culture is mistaken. This response

would put Rawls in a more disadvantaged position; many libertarians in fact find Rawls's egalitarianism too radical for a capitalist society. A fundamental reform of the basic structure and a strong egalitarian ethos are required for a Rawlsian well-ordered society.

Moreover, even if there is an overwhelming consensus on some political values, it does not entail that those values are desirable. Whether a value is justified does not depend on how many people actually accept it in a particular historical context. Rawls needs an *independent* argument to convince us that liberal values are the ideal moral basis of social cooperation. The public culture *per se* does not inform us which set of political ideas should be selected to govern the basic structure of society. The importance of certain values can only be judged in a wider moral context. Rawls, however, is reluctant to offer such an argument because he prefers presenting justice as fairness only as a freestanding political conception that does not involve any moral doctrines. But then, it is unclear why these values are overriding.

In addition, even if political values are proved to be significant, it does not automatically translate to an overriding motive on the part of citizens to act in accordance with them. For political values are just parts of a citizen's overall value system. "It may be overridden by citizens' comprehensive doctrines once all values are tallied up." (PL:386) For apart from political identity, citizens also affirm non-political identities found in the nonpublic life and associations they belong to. Rawls recognizes that citizens have other non-political aims and commitments which shape their ways of life and actions. "It can happen that in their personal affairs, or in the internal life of associations, citizens may regard

their final ends and attachments very differently from the way the political conception supposes." (PL:31) That being said, there must be a kind of unity between two identities grounded on one's comprehensive doctrine, or he would face serious internal conflict and become disorientated. The priority of political values would collapse accordingly. To avoid this, citizens must be able to perceive the political conception of justice as in some manner positively related to their comprehensive views. This is what Rawls calls the second complementary argument for political liberalism. He says:

> The history of religion and philosophy show that there are many reasonable ways in which the wider realm of values can be understood so as to be either congruent with, or supportive of, or else not in conflict with, the values appropriate to the special domain of the political as specified by a political conception of justice. (PL:140)

If this argument stands, a political conception of justice will be the focus of an overlapping consensus. Citizens will individually decide for themselves in what way the political conception of justice is related to their own comprehensive views. They comply with the requirement of justice for different reasons. The political conception gives no guidance in this regard; citizens' comprehensive doctrines are their guidance. Rawls believes that this more tolerant and pluralistic approach will have a better chance to win citizens' willing support of his political conception of justice. I will examine the plausibility of this approach in the next section.

## 3 Two Model Cases for a Consensus

It is now clear that the possibility of political liberalism depends on whether

justice as fairness could be widely accepted by reasonable comprehensive doctrines. Rawls's task is to demonstrate how the political conception is congruent, compatible, or at least not in conflict with citizens' conceptions of the good so that the problem of priority can be settled. In the following discussion I shall use two model cases to demonstrate that Rawls's argument for an overlapping consensus is unsuccessful.

To begin with, we should be reminded that Rawls's idea of an overlapping consensus is different from a modus vivendi. It has three distinct features. First, the object of consensus is itself a moral conception. Second, citizens are presumed to affirm the political conception on moral grounds. Finally, citizens are moved by the effective sense of justice rather than self-interest. In short, the consensus is not a consequence of a contingent balance of power. Rather, what it realizes is stability for the right reason. This ensures that "those who affirm the various views supporting the political conception will not withdraw their support of it should the relative strength of their view in society increase and eventually become dominant." (PL:148) When it comes to an overlapping consensus, the supporting reasons for the political conception are *moral* in nature.

In *Political Liberalism*, Rawls proposes several model cases to illustrate how a political conception is supported by reasonable comprehensive doctrines in different ways. They include a religious doctrine, Mill or Kant's liberal moral doctrine, utilitarianism, and a pluralist account of the realm of values. (PL:145,169-70) Rawls contends that all of them would accept justice as fairness based on the totality of reasons specified within their comprehensive doctrines. Their acceptance depends on two conditions. First, each doctrine will develop its

own reasons to support the political conception. Second, these reasons must be shown to be regulative in a citizen's motivational system. Below I will use Kantian liberalism and utilitarianism as examples to assess whether these requirements can be met.[5]

Rawls's account of the internal connection between Kant's moral philosophy and his political conception of justice is straightforward. He says:

> The first was of Kant's moral philosophy with its ideal of autonomy. From within his view, or within a view sufficiently similar to it, the political conception with its principles of justice and their appropriate priority, can, let us say, be derived. The reasons for taking the basic structure of society as the primary subject are likewise derivable. Here the relation is *deductive*, even though the argument can hardly be set out very rigorously. (PL:169, my emphasis)

Rawls presumes that if a person believes in a Kantian ideal of autonomy, he will deductively accept justice as fairness as a political conception. Among the four model cases, this comprehensive doctrine shows the strongest inclination to political liberalism. Since the principles are directly derived from the value of autonomy, Kantian citizens have sufficient motive to comply with the requirement of justice. However, the seemingly deductive relation between autonomy and Rawls's substantive principles is not as self-evident as Rawls supposes. A Kantian may not accept Rawls's principles as the best expression of moral autonomy. Nozick, for instance, ascribes the moral ground of libertarian side constraint to

---

[5] I focus on these two cases because they are regarded as competing and incompatible conceptions of justice in *A Theory of Justice*. It is particularly worthwhile to see how they can become members of an overlapping consensus in *Political Liberalism*.

Kant's moral doctrine as well. As he says, "side constraints upon action reflect the underlying Kantian principle that individuals are ends and not merely means; they may not be sacrificed or used for the achieving of other ends without their consent."[6] Autonomy demands respect for a person's self-ownership. Without his consent, no one has a right to take away his labour and property. Taking Kant's ideal of autonomy seriously requires a libertarian entitlement theory rather than a liberal redistributive scheme.

Rawls may argue that Nozick's interpretation of Kant is flawed. It is normal that different theorists have disagreements about the proper political implication of autonomy. The real problem, however, is that from the point of view of political liberalism Rawls cannot put such a challenge to Nozick. For following the logic of an overlapping consensus, it is left to each citizen to decide individually how the claims of political justice are embedded into the comprehensive doctrines he affirms. Having recognized the burdens of judgment, reasonable citizens must accept the fact that "many of our most important judgments are made under conditions where it is not to be expected that conscientious persons with full powers of reason, even after free discussion, will all arrive at the same conclusion." (PL:58) The burdens of judgment set a constraint on public discussion about the truth or reasonableness of a conception of the good. Nor can the freestanding political conception provide any external guidance on people's reasoning in this regard because "the guidance belongs to citizens' comprehensive doctrines." (PL:387)

---

[6] Nozick, *Anarchy, State, and Utopia* (Blackwell, 1974), pp.30-31.

One may counter that the Nozickean doctrine is so unreasonable that it should be excluded from the group of consensus. This is implausible because Rawls's definition of reasonable doctrine is deliberately loose. It only requires that the doctrine concerned is an exercise of theoretical and practical reason, and normally belongs to a tradition of thought and doctrine. Otherwise, the account "runs the danger of being arbitrary and exclusive." (PL:59) Moreover, a Nozickean could share the general idea of political liberalism. What he disagrees with Rawls about, rather, is that his commitment to autonomy would lead him to adopt libertarian principles. In this case, there is no more room for Rawls to maintain that his egalitarian distributive principles would necessarily be deduced from Kant's moral philosophy. This shows that even in the strongest case, the possibility of a consensus on Rawls's principles is indeterminate. In Rawls's own words, "whether justice as fairness (or some similar view) can gain the support of an overlapping consensus so defined is a speculative question." (PL:15) It must be speculative because the result is not determined by Rawls, but by numerous reasonable doctrines from their own point of view.

We can now turn to examine the relation between utilitarianism and justice as fairness, another model case presented by Rawls. This is much more controversial because Rawls has always presented his theory of justice as an alternative systematic account of justice to utilitarianism. But all of a sudden Rawls suggests that utilitarianism as a comprehensive moral theory would have sufficient reason to take justice as fairness as the most reasonable political conception for the basic structure. As Rawls's explanation goes:

> This utilitarianism supports the political conception for such reasons as our limited knowledge of social institutions generally

> and on our knowledge about ongoing circumstances...These and other reasons may lead the utilitarian to think a political conception of justice liberal in content a satisfactory, perhaps even the best, *workable approximation* to what the principle of utility, all things tallied up, would require. (PL:170, emphasis added)

The main reason for utilitarianism to adopt political liberalism is that it is the most effective means to maximize utility. This is a serious claim. If valid, there will be no conflict between two theories as Rawls presents in *A Theory of Justice*. Unfortunately, Rawls gives no further support for this claim. He simply supposes that justice as fairness is the most workable arrangement to meet the principle of utility. It is however unclear why this must be so. In principle, utilitarianism may favour a diversity of social institutions depending on their contribution to the greatest net balance of utility under particular social conditions. Thus, apart from Rawls's principles of justice, other competing non-liberal political conceptions of justice can make similar claims as well; which one is more plausible depends on empirical calculation.

But if Rawls's previous argument against utilitarianism is unchanged, we have reasons to doubt how it can be included in an overlapping consensus. Remember that the fundamental organizing idea of justice as fairness is of society being a fair system of cooperation for reciprocity between free and equal persons. The original position from which Rawls's political principles are derived is said to be designed in accordance with this idea. This implies that if utilitarianism accepts the political principles, it must accept Rawls's account of society as well. Yet in making a comparison between justice as fairness and utilitarianism, Rawls tells us that:

> Implicit in the contrasts between classical utilitarianism and justice as fairness is a difference in the underlying conception of society. In the one we think of a well-ordered society as a scheme of cooperation for reciprocal advantage regulated by principles which persons would choose in an initial situation that is fair, in the other as the efficient administration of social resources to maximize the satisfaction of the system of desire constructed by the impartial spectator from the many individual systems of desires accepted as given. (TJ:33/29-30 rev.)

We can note that Rawls's conception of society is fundamentally different from that of utilitarianism. Moreover, while justice as fairness takes seriously the plurality of distinct persons with separate system of ends, utilitarianism views the principle for a society as an extension of the principle of choice for one man. As a deontological theory, justice as fairness holds that the concept of right is prior to that of the good; but as a teleological doctrine, utilitarianism holds that "the satisfaction of any desire has some value in itself which must be taken into account in deciding what is right." (TJ:30/27 rev.) Finally, justice as fairness "does not allow that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many." (TJ:4/3 rev.) Since justice as fairness conflicts with utilitarianism in so many ways, it is impossible that a utilitarian would endorse Rawls's arguments for the rejection of utilitarianism while continuing to uphold utilitarian beliefs in their non-political domain. After all, all the disputes between justice as fairness and utilitarianism will appear misguided if Rawls's argument is right.

Furthermore, even if it is empirically proved that there is great social utility in accepting Rawls's political conception of justice, what it achieves would likely be stability for the wrong reason. We know that moral stability is secured by an

effective sense of justice specified by justice as fairness. Citizens are expected to act on moral reasons. In utilitarianism, however, the primary motive to abide by Rawls's political conception is simply the belief that it can best promote total utility. The liberal sense of justice does not play any role in determining utilitarians' action. If their limited knowledge of social institutions and knowledge about ongoing circumstances change, they have every reason to withdraw their support of justice as fairness for another institutional arrangement. Here there is a motive gap between utilitarianism and the liberal political conception. A faithful utilitarian will not have the moral motive required by political liberalism. His acceptance of the political conception is at most a modus vivendi.

Rawls dismisses this criticism. He insists that his conception of justice is based on moral grounds. They include "conceptions of society and of citizens as persons, as well as principles of justice, and an account of the political virtues through which those principles are embodied in human character and expressed in public life". (PL:147) Since these political ideas are drawn from the public culture and shared by reasonable comprehensive doctrines, utilitarianism as one of such doctrines would then give sufficient moral motive to respect the liberal political conception. Nevertheless, it should be noted that these political reasons to support justice as fairness are different from those stemming from one's comprehensive view. They represent two distinct types of justification. I follow Scheffler to call them *political* and *non-political* arguments:

> A political argument for a conception of justice would be one that appealed to ideas implicit in the public political culture, whereas a non-political argument, say, would be one that appealed to a comprehensive moral doctrine. Thus one and the same conception of justice might in principle be supported by

> arguments of either type. Rawls might then be interpreted as asserting not that his conception of justice is a political conception but, rather, that his arguments for that conception are political arguments.[7]

Scheffler's idea is that the same conception of justice can be argued for from two different perspectives. A utilitarian can either regard Rawls's conception of justice as a constitutive part of his comprehensive moral theory, or as a freestanding conception justified by the shared political values. In principle, the conception can be backed by either argument, each having its distinct normative source and motivational power. But I disagree with Scheffler that Rawls would regard these two arguments as equally important and independent of each other. Recall that the purpose of presenting justice as fairness as a political conception is to reach an overlapping consensus. Only when there is an overlapping consensus, according to Rawls, can the political conception be publicly justified. (PL:388) Therefore, the political argument, or *pro tanto* justification, is insufficient to vindicate stability for the right reason.

If the above analysis is correct, Rawls's response to the charge of modus vivendi misses the point. Even though his conception of justice is justified by a political argument, it may still be rejected by a comprehensive doctrine. This explains why utilitarians would abandon justice as fairness if they find better alternatives to promote the greatest balance of satisfaction. To avoid this predicament, one solution is to set a constraint on citizens' choice of the mode of justification. It can be stipulated, for instance, that in dealing with political

---

[7] Samuel Scheffler, *Boundaries and Allegiances* (New York: Oxford University Press, 2001), p.139.

questions of constitutional essentials and basic justice, citizens should only appeal to the political argument to settle their dispute. This is the position that Rawls holds when he discusses the idea of public reason. He suggests that on fundamental political questions citizens must honour the limit of public reason. The content of public reason consists of political principles and guidelines of inquiry that specify ways of reasoning relevant for political questions. It is a duty of civility not to use non-public reasons, namely reasons deriving from citizens' comprehensive views, in discussing and voting on the most fundamental questions. Furthermore, Rawls contends that "citizens affirm the ideal of public reason, not as a result of political compromise, as in a modus vivendi, but from within their own reasonable doctrines." (PL:218)

Rawls's idea of public reason raises a number of questions. For one thing, if citizens are all willing to set aside their comprehensive views and appeal to the same set of political values to settle the questions of basic justice, then the idea of an overlapping consensus is redundant. For the political argument alone is sufficient to confer priority to political values over non-political ones. This, however, is question-begging. What we have been asking all along is how it can be either reasonable or rational for citizens to be motivated to act on political values given the fact of reasonable pluralism. The quest for consensus is necessary because Rawls acknowledges that people take their non-political beliefs and commitments seriously in their practical reasoning. That is why the non-political argument is needed. Once this is granted, as Scheffler notes, "any requirement that the participants in an overlapping consensus must view the conception of justice

as political would appear to be incongruous with the motivation for introducing the idea of such a consensus in the first place."[8]

Furthermore, an inconsistency exists between the idea of public reason and the idea of an overlapping consensus. The former requires that one should only appeal to political values while the latter allows a diversity of reasons drawn from comprehensive doctrines. With a closer look at Rawls's account, we will find that this is a misunderstanding. Rawls's point is that the limit of public reason is the result of an overlapping consensus. Only when the political conception is shown to be the focus of an overlapping consensus will citizens have sufficient reason to honour public reason. Rawls says:

> Political liberalism relies on the conjecture that the basic rights and duties and values in question have sufficient weight so that the limits of public reason are justified by the overall assessments of reasonable comprehensive doctrines once those doctrines have adapted to the conception of justice itself. (PL:219)

Rawls believes that an overlapping consensus entails the limits of public reason. When a conception of justice is shown to be accepted by a plurality of reasonable comprehensive doctrines, citizens will have sufficient motive to limit themselves to the use of political values in discussing fundamental political matters. This account gets Rawls into more troubles though. If an overlapping consensus is the precondition of public reason, it is then illogical for Rawls to appeal to the limits of public reason to justify the possibility of including utilitarianism as a member of the consensus. Their relation cannot be reversed.

---

[8] Scheffler, *Boundaries and Allegiances*, p.141.

Further, it should be recalled that the non-political argument is not only legitimate, but also essential to the very idea of an overlapping consensus. Therefore, my contention that utilitarianism would at most adopt a liberal conception of justice as a modus vivendi remains intact.

## 4 The Limits of the Burdens of Judgment

I have examined two model cases and cast my doubt upon the possibility of an overlapping consensus. We note that the difficulty of Rawls's project lies in the division between the political and non-political spheres. The aim of this division is to give room for a convergence of reasonable comprehensive doctrines on a freestanding political conception. Each comprehensive doctrine is encouraged to develop its own reason to justify the political conception. However, once the plurality of justificatory reasons is granted, there will be no way for Rawls to assure a consensus for the right reason.

Rawls, however, would complain that my critique has overlooked an important qualification of his project, namely all members of an overlapping consensus are supposed to be reasonable persons who are willing to recognize the burdens of judgement in political justification. The burdens of judgement warrant that a liberal conception of justice will be the *only* acceptable choice available for reasonable citizens who hold reasonable comprehensive doctrines. It sets a strong constraint on citizens' choice of principles. Rawls therefore need not worry too much about my concern that citizens may not give priority to political reasons from their own point of view. For their point of view has been limited in the first place. We can note that the strategy of this argument is similar to that of rational

choice in the original position. In this section, I will argue that this account of reasonableness is inconsistent with the primary spirit of political liberalism.[9]

Rawls's main ideas are as follows. To begin with, he assumes that all participants of cooperation are reasonable persons. Persons are reasonable in two aspects. The first is the willingness to propose fair terms of cooperation and to abide by them. The second is the willingness to recognize the burdens of judgement and to accept their consequences. The burdens of judgement refer to those sources that result in reasonable disagreement about many of our most important judgments. Rawls regards this as a normal result of the exercise of human reason within a liberal democracy. It is a permanent feature of the public culture of democracy. (PL:36) The burden of judgement entails that a public conception of justice can never be justified by any comprehensive religious, philosophical, or moral doctrines. For this will violate the liberal principle of legitimacy according to which "our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse." (PL:137) Thus, it is unreasonable for citizens to use political power to repress any reasonable comprehensive views that are different from their own. Having recognized the burdens of judgment and the principle of legitimacy, it naturally leads to the conclusion that only a political conception of justice as a freestanding view presupposing no particular comprehensive doctrines will be accepted by reasonable citizens. As Rawls puts it, "reasonable persons see that the burdens of judgement set limits on what can be reasonably justified to others, and so they

---

[9] I am indebted to John Charvet for his helpful advice on the argument of this section.

endorse some form of liberty of conscience and freedom of thought." (PL:61) Therefore, reasonable citizens must accept the burdens of judgement which in turn commits them to a liberal political conception of justice.

We can now start to evaluate this central argument of political liberalism. First of all, the claim of the burdens of judgement may backfire on Rawls's own conception of justice. The reason is simple. If it is true that reasonable people will have great difficulty in arriving at the same conclusion about many of their beliefs and values in the non-political sphere even after conscientious and free discussion, why does a similar situation not apply to political values? There seems no apparent reason for Rawls to draw such a distinction and give privilege to moral values. Even if we agree that political values are implicit in the public political culture, it does not follow that these values will be generally acceptable to reasonable people. So the burdens of judgement will not only exclude the possibility of grounding justice on a moral doctrine, but also undermine the whole scheme of political liberalism. However, this claim is unreasonably strong. Moral experience tells us that in many cases we can reach informed and reasoned judgment on moral and political issues. It is too early for Rawls to divide human values into two spheres, and to assign a different status to them.

My second challenge is that the idea of reasonableness imposes some unreasonable demands on citizens. It should be noted that the claim of burdens of judgement is a self-standing argument for justification of political liberalism. It is not derived from a person's comprehensive doctrine. For example, a religious believer will maintain that what he believes is true. Even in face of the challenge of pluralism, he could reject other religions as false. It is almost impossible for a

devout believer to accept that other competing religions are equally true or reasonable. Doing this would fundamentally undermine his belief. But for the sake of political justice, Rawls demands that whatever a person believes in, it must be subject to the burdens of judgment. That means the person has no ground to appeal to his belief to justify any principles of justice. He should recognize that "this is a claim that all equally could make." (PL:61) This is hardly acceptable for persons who have deep belief in their moral and religious doctrines. They may ask why they should set aside their claim that "the duty to religious and divine law being absolute, no understanding among persons of different faiths is permissible from a religious point of view." (TJ:208/182 rev.)

Rawls actually admits that people have often acted in accordance with this doctrine. However, the acceptance of the burdens of judgment, Rawls argues, will urge them to realize that reasonable people do not all affirm the same comprehensive doctrine. As a result, "we recognize that our own doctrine has, and can have, for people generally, no special claims on them beyond their own view of its merits. Others who affirm doctrines different from ours are, we grant, reasonable also, and certainly not unreasonable." (PL:60) People can continue to believe their doctrines to be true; meanwhile, they should accept that the truth of their belief does not apply to other reasonable people. If they insist on doing so, they are unreasonable. But how could a person consistently uphold the truth of his comprehensive view and accept the consequence of the burdens of judgment in the meantime? The latter seems to imply that a reasonable person must hold a position of moderate scepticism which, according to Barry, believes that "no

conception of the good can justifiably be held with a degree of certainty that warrants its imposition on those who reject it."[10] This moderate view does not reduce normative statements to expression of personal preference and subjective feelings. Its central idea is to express an attitude of uncertainty towards the truth of any conceptions of the good attested by experience. I believe that Rawls's account of burdens of judgment actually amounts to this sceptical position. It requires people to realize and accept that no matter how true or reasonable their conceptions of the good are viewed from their own point of view, they can always be doubted and reasonably rejected by other people. When moderate scepticism is combined with another aspect of reasonableness, namely that people have a desire to propose principles which others could not reasonably reject, a freestanding political conception will be the only acceptable candidate for fair cooperation.

Rawls does not accept this interpretation of the burdens of judgement. For if political liberalism is grounded on a sceptical argument about conceptions of the good, the idea of an overlapping consensus would fail. He thus says:

> Political liberalism does not question that many political and moral judgments of certain specified kinds are correct and it views many of them as reasonable. Nor does it question the possible truth of affirmations of faith. Above all, it does not argue that we should be hesitant and uncertain, much less skeptical, about our own beliefs. (PL:63)

Rawls here has adopted a method of avoidance to respond to the challenge. Political liberalism deliberately avoids making any judgment about the truth of citizens' beliefs. It stands outside and allows citizens to decide for themselves the

---

[10] Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), p.169.

epistemological status of their conceptions of the good. Political liberalism is neutral and practical. It simply recognizes the practical impossibility of reaching any reasonable and workable political principles based on the truth of conceptions of the good. This explanation, however, is not convincing. We note that the burdens of judgement set normative and epistemological constraints on citizens' practical reasoning. Citizens are required to view their most fundamental beliefs and commitments from a sceptical point of view. If they refuse to do so, they will be regarded as unreasonable and excluded from an overlapping consensus. It is thus misleading for Rawls to say that the burdens of judgment are simply objective characterisations of the fact of reasonable disagreement. If my argument is plausible, scepticism will substantially undermine Rawls's claim that citizens will converge on a liberal conception of justice from their own point of view without having to make any change of their conceptions of the good.

One may ask if the account of reasonableness has already determined some form of liberal political conception of justice, why does Rawls need the second stage to handle the problem of an overlapping consensus? This is because citizens have to harmonize their acceptance of the political conception with their non-political values. The reasonable alone is insufficient to warrant the priority of political values. The political conception must be shown acceptable within a citizen's comprehensive view. This concern brings us back to my previous critique of Rawls's two arguments for the possibility of political liberalism, namely 1) the greatness of political values and 2) the existence of many reasonable ways of making the political conception of justice congruent with comprehensive doctrines. In my previous discussion, I have used Kantianism and utilitarianism as examples to demonstrate the difficulty of consensus. If we think further, we will

find that those non-liberal comprehensive doctrines, such as Platonism, Islamism and Thomism, will have even greater difficulty in accommodating liberalism in their worldviews. They have to liberalize themselves in both stages of justification. At the first stage, they are required to accept the consequence of the burdens of judgment; at the second stage, they are required to accept the priority of political liberal values over their non-political non-liberal values. Unfortunately, Rawls offers no answer why these non-liberal doctrines are willing to liberalize themselves in the way political liberalism suggests.

To conclude, this chapter has examined Rawls's three essential arguments for an overlapping consensus. I have shown that all of them are insufficient to establish the priority of political values through the support of an overlapping consensus. In other words, political liberalism fails to provide a better alternative to resolve the problem of stability. In this case, we have good reasons to search for other possibilities. This is what I attempt to do in the last chapter of this thesis.

# CHAPTER 6

# POTENTIAL CONGRUENCE

My thesis has so far been devoted to two fundamental questions, namely to what extent Rawls's notion of stability can be justified, and if it can, whether justice as fairness can offer a satisfactory answer to it. For the first question, I affirm Rawls's claim that moral stability in terms of the motivational overridingness of a sense of justice is essential to the justifiability of justice as fairness. For the second question, I contend that the congruence argument of *A Theory of Justice* and the idea of an overlapping consensus of *Political Liberalism* have both failed. The former does not stand because the Kantian teleological account of human nature is incompatible with prudential rationality and reasonable pluralism; whereas the latter cannot adequately show how reasonable comprehensive doctrines would honour the priority of political values from their own point of view. In this light, the remaining question is whether it is possible to find another way to vindicate the stability of justice as fairness. This last chapter attempts to present such an alternative argument for the motivational priority of justice.

I will call this approach *Potential Congruence.* It aims to vindicate the idea that it is rational for a person to give precedence to morality over narrow self-interest because leading a just life can be presented as a higher-order regulative good in one's rational plan of life under favourable conditions. Put it another way, there can be sufficient reasons for a person to accept morality as a hierarchically superior value, and to form and pursue his conception of the good as a whole subject to the requirement of morality. The right is not alien to the

good. Rather, acting justly itself is essential to one's well-being. This is not a new argument. Its spirit is not much different from Rawls's idea of congruence. Similar to Rawls, potential congruence strives to attest the possibility that the desire to act justly and the desire to lead a good life are practically speaking the same desire. What is different is that it does not appeal to a Kantian metaphysical account of human essence. It no longer holds that acting justly is the only way to express our human nature as free and equal rational being. It recognizes that this claim is too strong to be acceptable in a society of reasonable pluralism.

Potential congruence is based on a conception of human reason and interest that can be accepted by rational and reasonable people. It is a result of practical reasoning supported by substantive arguments rather than a forgone conclusion guaranteed by definitional or metaphysical truth. Besides, I am not making a general claim that all conceptions of justice can achieve potential congruence. This is implausible because different conceptions make different demands on agents who in turn have different reactions to these claims. My discussion is more specific. I will first lay out some general conditions for congruence and then explore whether Rawls's principles of justice can be a focus of congruence. Having said that, what this chapter presents is still a skeleton outline for this big issue.

This chapter consists of five sections. In the first two sections, I will discuss two pre-conditions for potential congruence. They are the unity of practical reasoning and the pervasiveness of moral feelings. After that, I proceed to examine whether it will be rational for an agent to give priority to justice as fairness. In Section 3, I discuss Rawls's idea of society as a fair cooperation for

reciprocity which provides a foundational framework for Rawls's two principles of justice. I then focus on the first principle of equal basic liberties in Section 4. Through Rawls's arguments for the priority of basic liberties, I show that it is rational to affirm the good of the sense of justice specified by this principle. The last section turns to examine the difference principle. I contend that the economic incentive argument underlying the difference principle is inconsistent with moral equality. I then propose the idea of the "modification of impersonal morality" as an alternative to justify the principle, which may therefore have a better chance to achieve potential congruence.

## 1 The Unity of Practical Reasoning

I believe that the claim of potential congruence depends on two conditions, namely the unity of practical reasoning and the pervasiveness of moral sentiment. They provide background support for my subsequent claim that congruence is a viable and desirable option for moral stability. I will explore the meaning and implication of these two claims in this and the next sections.

The unity of practical reasoning holds that a rational agent normally has a fundamental interest in grounding his reason for action on a unified and coherent value system. This view presupposes that we envisage each human life as a whole. We do not view our life as a series of unconnected episodes, or think about our action as a sequence of unrelated and independent parts. For this would disintegrate our life and make it unintelligible. To understand someone's action, we must have an accurate understanding of his intentions in a particular social and

historical setting.[1] Moreover, as purposive beings, we desire to lead a meaningful and valuable life. It is our desire to justify to ourselves - and more often to others - that our life is worth living. We are not indifferent to our choice because our lives are irreplaceable and enormously important viewed from inside. Therefore, we need a normative framework to help us make decisions. This framework will provide explanation and justification for our actions, and give our life a unitary shape. Only so can the diversity of intentions, desires, ideals and projects form an ordered and coherent whole within our lives.

This does not mean that our life will never become fragmented and divided. On the contrary, as MacIntyre notes, one of the characteristic features of modernity is to partition human life into segments, each with its own norms and modes of behaviour. "So work is divided from leisure, private life from public, the corporate from the personal."[2] We have many roles to play. These roles may sometimes come into conflict with each other. I am not saying that rationality demands a quest for the unity of a life, though I tend to believe that more often than not rational actions presuppose such a normative framework.[3] I prefer a moderate view instead which emphasizes only a commonplace phenomenology that most people under normal circumstances have a fundamental desire to view their life as a continued and unified whole based on a normative scheme. Without such a scheme, we might not be able to tell ourselves or others why we have made such and such decisions and why those are meaningful and significant; we might

---

[1] MacIntyre has a good discussion about this issue. *After Virtue* (Notre Dame, Indiana: University of Notre Dame Press, 1984), 2nd edition, chap.15.

[2] MacIntyre, *After Virtue*, p.204.

[3] Rationality refers to means-end rationality. I am not able to deal with this complicated issue here. For more discussion, see Michael E. Bratman, *Intention, Plans and Practical Reason* (Cambridge, Mass: Harvard University Press, 1987)

not know where we were from, and where we are to go; we might also lack the resources and criteria to resolve conflicts of values and obligations. When a person's life is broken into many disconnected parts, he might probably become lost and disoriented. Therefore the quest for the unity of practical reasoning is entrenched in our everyday life. Everyday vocabulary such as "plan", "project", "scheme", "conception", "worldview" and the like reflects this very deep psychological need of human beings.

Rawls calls this normative framework a conception of the good, or a rational plan of life. It includes "conceptions of what is of value in human life, and ideals of personal character, as well as ideals of friendship and of familial and associational relationships, and much else that is to inform our conduct, and in the limit to our life as a whole." (PL:13) In this sense, a person may be viewed as a human life lived according to a conception of the good. "An individual says who he is by describing his purposes and causes, what he intends to do in his life." (TJ:408/358 rev.) Besides, Rawls regards each person having a plan of life as part of the subjective circumstances of justice. "These plans, or conceptions of the good, lead them to have different ends and purposes, and to make conflicting claims on the natural and social resources available." (TJ:127/110 rev.) Without this assumption, Rawls contends, "there would be no occasion for the virtue of justice" (PL:128/110 rev.) The good is conceptually prior to the right in this sense. Thus, the primary motive for a group of people to gather together and agree on a conception of justice is exactly that they can better advance their antecedent interests this way. Society is therefore a cooperative venture for mutual advantage, with the principle of the right defining the terms of cooperation.

For Rawls, a person's conception of the good constitutes the basis for the unity of practical reasoning. It offers a normative framework to integrate different parts of our life into a whole, and motivates us to act in certain way. This has great implications for the claim of potential congruence. If we agree that nothing can count as a reason for a rational agent unless it is capable of motivating him, then the source of motivation must be viewed as stemming from his conception of the good. It follows that if an agent is willing to give regulative priority to moral considerations over other desires, it is because he has instilled into his conception of the good a respect for morality. The sense of justice is conceived of occupying a regulative place in an agent's motive system which allows him to subject his overall conception of the good over a whole life to the constraints of justice specified by a political morality.[4] The desire to act justly is not a peculiar motive detached from a person's plans and projects. Rather, the priority of the sense of justice is justified by establishing its supreme status in one's rational plan of life.

We can note that Rawls's two kinds of argument for stability have presupposed the unity of practical reasoning. His account of congruence aims to show it is rational that "for those in a well-ordered society to affirm their sense of justice as regulative of their plan of life." (TJ:567/497 rev.) The central idea of an overlapping consensus is to allow citizens individually to decide how the values of the political domain are related to their comprehensive conceptions of the good. Although citizens do not hold the same comprehensive doctrine, Rawls hopes that citizens can accept the priority of justice from their own point of view. Ideally, "there are many reasonable ways in which the wider realm of values can be

---

[4] I am indebted to John Charvet for clarifying this point.

understood so as to be either congruent with, or supportive of, or else not in conflict with, the values appropriate to the special domain of the political as specified by a political conception of justice." (PL:140) Two arguments, though fundamentally different, basically share the same assumption that the desire to act justly must in some way be related to a person's conception of the good.

This assumption of practical reasoning, however, could sound disturbing as it potentially vitiates the authority of morality. It might be argued that if what one ought to do depends on whether that action can best realize his conception of the good, the normative force of moral reason will then be reduced to the mere satisfaction of self-interest. A person will have no reason to honour the demand of morality if he finds abiding by moral principle not conducive to his interests. The qualitative difference between morality and self-interest is apparently dissolved. This contradicts the very idea of moral overridingness, that is, the idea that moral reason should overrule concern for one's own good whenever the two sorts of consideration diverge. After all, if these two kinds of reasons are not conceptually different from each other, strictly speaking, the problem of overridingness would not arise in the first place.

This challenge, though serious, depicts a misunderstanding about the nature of potential congruence. First of all, we should note that when an agent's conception of the good is said to provide a basis for his practical reasoning, it does not mean that he is purely moved by self-interest to realize an egoistic goal. We should draw a distinction between interest in the self and interest of a self. While it is each rational individual's interest in regarding his conception of the good as worthy of recognition and realization, the content of the conception of the good is

not presumed to be egoistic or selfish. It all depends on what kind of ends the person is pursing. As Rawls puts it, "if wealth, position, and influence, and the accolades of social prestige, are a person's final purposes, then surely his conception of the good is egoistic. His dominant interests are in himself, not merely, as they must always be, interests of a self." (TJ:129/111 rev.)

A conception of the good is a formal idea designed for explaining the structure of practical reasoning. Its substantive content is filled in by agents themselves. It could encompass one's religious, metaphysical and moral beliefs, interests and desires, aims and ideals, projects and commitments, and personal and impersonal interests. This is similar to Williams's idea of "subjective motivational set" which may include:

> Such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent. Above all, there is of course no supposition that the desires or projects of an agent have to be egoistic; he will, one hopes, have non-egoistic projects of various kinds, and these equally can provide internal reasons for action.[5]

We thus see that although potential congruence has adopted a desire-based theory of practical reason, the term "desire" refers to a wide range of human dispositions. Commitment to moral principles could be an important element in one's conception of the good.[6] The unity of practical reasoning just stipulates that the reason to act justly must be situated in an agent's motivational set. It, however,

---

[5] Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), p.105.

[6] Rawls holds that his account of moral motivation can be viewed as belonging to a person's motivational set as Williams stipulates. See PL:85, footnote 33.

does not say that moral reason is only instrumental to the agent's narrowly defined self-interests. Impartial and partial concerns can co-exist in one's conception of the good. Of course whether a person is willing to give priority to moral reason depends upon a balance of motives.

It is therefore misleading to say that the claim of potential congruence will in effect reduce moral reason to the satisfaction of self-interest. The claim does not hold that other-regarding reason and self-regarding reason are identical. Suppose the former refers to the moral point of view which is identified with a distinctive set of impartial considerations, whereas the latter is only concerned with the realization of personal well-being,[7] potential congruence recognizes the potential conflict between these two distinct perspectives. For neither does it define morality in terms of the agent's interest, nor define the agent's good life in terms of moral life.[8] There is no conceptual connection between the dictates of morality and the pursuit of personal interest that could suppress all conflict between the two. The moral perspective has its independent status in one's conception of the good. What the claim of potential congruence aims to establish is that under favourable conditions a reasonable conception of justice can be shown to be congruent with a person's pursuit of well-being. As Scheffler aptly describes the idea, it strives to vindicate the claim that "moral norms should be capable of being integrated in a coherent and attractive way into the life of the individual agent."[9]

---

[7] What should be counted as a moral point of view is a complex issue that I cannot discuss in detail here. But it is widely accepted that it should involve an impartial attitude toward other people.

8 According to Nagel, the first position is held by Aristotle while the second by Plato. *The View from Nowhere* (New York: Oxford University Press, 1986), pp.195-96; also see Scheffler, *Human Morality* (New York: Oxford University Press, 1992), p.54.

9 Scheffler, *Human Morality*, p.102.

It is a consequence of substantive justification rather than that of conceptual guarantee.

It might be questioned if a majority of rational individuals, after deliberation, still insist that they have no motive to act justly, does this mean that the authority of morality will lose its ground? The answer depends on what reason they would give to support their claim. There are at least two possibilities. First, these people are purely egoists so that morality has little hold on them. For egoists, moral reason would never have independent normative force. It is merely a means to realize their self-interest. Self-interest would trump moral considerations should they come into conflict.

Those who attempt to dissuade egoists may argue that acting morally is always beneficial to one's personal interests. It is irrational for them not to do what morality requires. This claim is, however, too strong to be true. As I have shown in Chapter 4, prudential rationality does not have such power to vindicate the unexceptional congruence between justice and self-interest in a pluralistic society. We cannot provide a conclusive reason to convince an egoist to accept the overridingness of the sense of justice if he is not willing to commit to a moral point of view in the first place. In presenting his argument for congruence, Rawls acknowledges this point and stresses that we should not evaluate the goodness of the sense of justice from an egoistic viewpoint. Rather, "we are concerned with the goodness of the settled desire to take up the standpoint of justice." (TJ: 568/498 rev.) Members of a well-ordered society already possess a well-grounded moral sentiment. The question is whether it is rational to give a regulative priority to the sense of justice to overrule self-interest in case of conflict. It may be

complained that Rawls's assumption is too idealistic to be applicable to our society. I do not think this is the case. I believe that a wide range of moral sentiments is actually embedded in most people's lives, which in turn provides a solid basis for the possibility of congruence.

We can now turn to examine the second possible account of the failure of moral authority. According to this account, the failure results from excessively demanding moral principles rather than from egoism. The real problem is not that rational people lack a moral motive to do what justice requires; they are supposed to have a sense of justice. What they find unacceptable is that the moral constraint specified by a moral theory is too harsh to be compatible with their conception of the good. So it is a challenge to the motivational accessibility of a theory rather than the authority of morality in general.

The degree of a theory's demandingness is a function of a number of factors. According to Scheffler, two of them are especially important:

> One is the *extent* to which the theory's constraints are confining: that is, the extent to which they narrow the range of morally acceptable courses of action open to an agent. The other is the *cost* to the agent of satisfying the theory's requirements, which in turn is a function of such things as the degree of incompatibility, whether logical, physical, psychological, or practical, between what the theory requires the agent to do, and what it is in the agent's own interest to do.[10]

How to decide an optimal level of demandingness of a theory is an important issue that we cannot deal with here. It suffices to note that if a theory leaves

---

[10] Scheffler, *Human Morality*, p.98, emphasis added.

agents with little room to pursue their life projects, or demands a great deal of sacrifice of their interests to meet the requirement of justice, they may have a legitimate reason to reject the claim of overridingness. This is not because they are evil or self-centred. For they acknowledge the independent force of a moral standpoint and accept that morality can make claims on them. But it is unreasonable to expect that an impartial assessment of the value of a person's life should exhaust this very life. An impartial standpoint is only one of the many parts of our life. Moreover, one's own projects and interests can sometimes carry a disproportionate weight in determining what one may permissibly do. As rational autonomous beings, they have a fundamental interest in realising their conceptions of the good. A reasonable political morality should take this factor into account. If the moral demand of a theory causes strong tension in people's lives, we can foresee that the chance of attaining congruence would be slim. In short, the pursuit of moral stability warrants a reasonable constraint on the content of principles of justice.

## 2 The Pervasiveness of Moral Feelings

Discussion above shows that potential congruence depends on a balance of motives. On the one hand, members of a cooperative scheme should have a settled desire to act morally. They are not moved solely by self-interest. They are willing to discuss fair terms of cooperation with others from an impartial perspective and to abide by principles that they find morally acceptable. On the other hand, the political conception of justice must not set excessive constraints on agents' pursuit of the good life. It should give people sufficient autonomy to form and develop their projects within a moral framework. A conception of justice that satisfies

these conditions would then stand a good chance to vindicate the claim that observing the principles of justice is an essential good in our well-being.

This looks like the right direction to resolve the problem of stability. Nevertheless, one might argue that this proposal is unrealistic because people are egoistic by nature. This nature is further exaggerated in an individualistic capitalist society characterized by unconstrained quest for wealth and power. Egoism prevails over morality. Potential congruence is impossible in practice. Admittedly, if we are living in an egoist society, it is hard to justify the precedence of the sense of justice. It is also undeniable that the development of capitalism has substantially eroded people's moral concern about others. Our motives are fundamentally shaped and molded by the social and political system. Living in a competitive and individualistic market society makes it harder for people to accept the claims of liberal egalitarianism as a regulative good of their lives. I am well aware of these difficulties.

However, this does not mean that potential congruence is doomed to failure. First of all, following Rawls, what I am concerned with is the possibility of congruence in a well-ordered society in which its members have an effective sense of justice and the basic structure is satisfied with a public conception of justice. Secondly, I do not believe that our actions are solely moved by egoistic concerns. No matter how imperfect our society is, we are living in an ethical community. Our mental and social lives are fundamentally shaped by moral beliefs and moral sentiments since we were born. Moral concerns play an essential role in forming our conceptions of the good and determining our actions. I call this phenomenon the pervasiveness of moral sentiments, the second

condition for potential congruence.[11] The term "moral sentiments" generally refers to those moral attitudes, feelings, or emotions that have a significant and enduring place in a person's life such as the sense of justice and the love of mankind.[12]

As Rawls points out, a variety of moral emotions is interwoven throughout the fabric of human personalities and social life. These powerful feelings cannot be properly understood without presupposing moral beliefs and moral principles. Guilt, shame, indignation, remorse and resentment fall into this category. Take guilt as an example. When we ask a person why he feels guilty, it is not enough for him to describe his feeling as a mixture of fear, anxiety, and regret. Nor can it be explained by expected punishment. Rather, the explanation must invoke a moral concept and its associated principles. The person's experience of guilt must result from doing something morally wrong. For example, he knows that he has taken more than his fair share in a distribution as defined by a conception of justice. Similarly, a person feels ashamed because he has failed to live up to virtues defined by a conception of moral worth. Thus, guilt and shame reflects our concern with others and with our own good. "In general, guilt, resentment, and indignation invoke the concept of right, whereas shame, contempt, and derision appeal to the concept of goodness." (TJ:484/423 rev.) This implies that we could not experience these important emotions if we do not have pre-existing moral beliefs and principles; to have those beliefs, we must be moral beings in the first

---

[11] On this issue, I am greatly indebted to Scheffler's discussion. He calls this phenomenon the "resonance of morality." See *Human Morality*, pp.68-70.

[12] It should be noted that Rawls has drawn a subtle distinction between moral sentiment, moral attitude and moral feeling. Since his classification will not affect my argument, I will use them interchangeably to refer to the same meaning. (TJ: 479-80/420 rev.)

place.

If the aforesaid is correct, then an egoist who would never act from a sense of justice is incapable of experiencing these moral feelings. A person cannot feel guilty if he does not have a conception of right and fairness. Or as Scheffler describes, "one can feel angry at being ill treated without having any moral beliefs, but one cannot resent the ill treatment unless one believes that it was wrong or unjustified or unfair."[13] This is because resentment is a moral attitude whereas anger is a natural feeling.[14] Bernard Williams makes a similar observation when he remarks that it would be perfectly consistent for an amoralist to object to other people treating him in the same way as he treats them so long as "his objecting consists just in such things as his not liking it and fighting back. What he cannot consistently do is to resent it or disapprove of it, for these are attitudes within the moral system."[15] A person who has no moral attitudes and who therefore never acts out of justice would be bound to strike us as humanly incomplete. As Rawls puts it, "one who lacks a sense of justice lacks certain fundamental attitudes and capacities included under the notion of humanity." (TJ:488/428 rev.) So being an egoist is not only undesirable, but practically impossible in an ethical community. When a person grows up in a well-ordered society, he will naturally develop a diversity of natural and moral sentiments through family, associations and public institutions. Once acquired, it may be difficult for the person to give them up at will. Moral concerns are deeply embedded in people's practical reasoning. They

---

[13] Scheffler, *Human Morality*, p.68.

[14] A stimulating discussion about resentment can be found in P.F. Strawson, "Freedom and Resentment," in his *Freedom and Resentment and Others Essays* (London: Menthuen, 1974).

[15] Bernard Williams, *Morality* (Cambridge: Cambridge University Press, 1976), p.5.

are a normal part of human life. Those people who have reservations about congruence normally believe that self-interest gives an agent the most primary and strongest reason to act. There is an unbridgeable gap between living rightly and living well. However, if the claim of the pervasiveness of moral feeling is correct, egoism does not set a real challenge to the possibility of congruence.

So far we have discussed the claims of the unity of practical reasoning and the pervasiveness of moral sentiments, both of which provide support for the possibility of congruence of the right and good. As Nagel aptly observes, "if it is the function of an ethical theory to identify both the moral life and the good life, and to reveal the reasons we have to lead each of them, then a theory that allows them to diverge will be claiming something that is hard to accept, given the importance of each of these ideals."[16] Nevertheless, this by no means implies that the society we find ourselves in is completely just. Nor does it claim that each individual has equally acquired an effective sense of justice to honour the requirement of justice. What the claim of the pervasiveness of moral sentiments establishes is that "moral concerns resonate throughout the web of human social relations."[17] It denies that concerns of self-interest constitute the sole motive in practical reasoning. But this phenomenon alone is far from enough to prove the priority of the sense of justice specified by a political theory. To justify congruence, we need further argument to help us judge whether justice as fairness can be the focus of congruence between the right and the good. This is what I am going to examine in the rest of this chapter.

---

[16] Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986), p.205.

[17] Scheffler, *Human Morality*, p.78.

## 3 The Value of Social Cooperation

In this and the following sections, I will attempt to vindicate the potential congruence of justice as fairness. My discussion will focus on the following question: is it rational for an agent to accept the priority of Rawls's principles of justice in social cooperation? If the answer is positive, then the question of overridingness of the sense of justice will be settled. This section will examine the value of social cooperation; the next section will discuss the good of basic liberties, and the last section will evaluate the difference principle. But I cannot here present the argument for such congruence in a rigorous and comprehensive manner, but shall merely indicate the direction that we should proceed in.

Some qualifications should be first noted. First, rationality refers to prudential rationality, according to which an action is rational if it can best promote an agent's informed interests, or rational conception of the good.[18] Interests are understood in its broadest sense including one's ends, projects, ideals, and impartial concerns for others. Second, my discussion is located in what Rawls calls the second stage where the principles of justice have been worked out already. Participants in cooperation know the justificatory reasons for Rawls's principles of justice and their particular conceptions of the good. They also approve the fundamental idea of society as a fair system of cooperation between free and equal citizens. The challenge of congruence is to show how the requirement of justice and citizens' conceptions of the good can be harmonized,

---

[18] On the definition of the rational conception of the good, see *TJ*, Chapter VII. The conception of prudential rationality should be distinguished from that of instrumental rationality which is defined by the most effective means to satisfy an agent's *existing* desires.

and the priority of the former can be firmly established by affirming the good of the sense of justice. Finally, as already argued above, I assume that people have a settled and effective sense of justice. They are not moved by self-interest though they have a fundamental interest in advancing their own conception of the good. In short, potential congruence does not attempt to justify liberal egalitarianism to egoists. It asks another question: suppose citizens have a moral motive, why should they affirm it as a fundamental and regulative value in their conceptions of the good?

With the above qualifications in mind, our discussion will start with Rawls's conception of social cooperation. Rawls holds that cooperation involves three basic elements. First, cooperation requires a set of publicly recognized rules and procedures to determine participants' rights and duties, and benefits and burdens. Second, these rules specify fair terms of cooperation for reciprocity which each participant may reasonably be expected to accept. Reciprocity means that all participants can benefit in an appropriate way as assessed by a suitable benchmark of equality. Third, cooperation requires an idea of each participant's rational good which specifies what they aim to achieve through that undertaking. Rawls refers to this as one's conception of the good. (PL:16)

This conception of social cooperation is a very fundamental idea of Rawls's whole enterprise. It conveys at least three important messages. First, people recognize that social cooperation is a fundamental common good for every participant. "There is an identity of interests since social cooperation makes possible a better life for all than any would have if each were to live solely by his own efforts."(TJ:4/4 rev.) Cooperation not only improves our conditions of living,

but also realizes many valuable human capacities through associations and social unions. We can enjoy and appreciate one another's excellences and individuality by participating in different types of collective activities. Besides, as a matter of fact, we are born into society. For most of us, there is no entry or exit except by birth and death. "There is no alternative to social cooperation except unwilling and resentful compliance, or resistance and civil war." (PL:301) Thus, it is reasonable to expect that persons have a strong reason to participate in social cooperation.

Second, social cooperation requires a set of rules to regulate and coordinate participants' behaviour. These rules must be fair and acceptable to every equal participant so that we are willing to cooperate in good faith with all members over a complete life. To realize this ideal, the rules must enjoy an authoritative and overriding status. They should not be subject to people's relative bargaining strength. For this would make cooperation unstable. What they want are moral terms of cooperation based on mutual respect. As Charvet remarks, "cooperation on moral terms has the advantage that the terms—being necessarily authoritative for everyone's self-interest—cannot be subject to challenge whatever changes occur in the relative bargaining position of the cooperators."[19] Thus, participants have a higher-order interest in setting up an authoritative framework within which they can pursue their conception of the good. The overridingness of norms is a prerequisite for fair cooperation for reciprocity.

Third, if persons have a regulative desire to participate in cooperation, it is

---

[19] John Charvet, *The Idea of an Ethical Community* (Ithaca & London: Cornell University Press, 1995), p.168.

reasonable to assume that they have a corresponding desire to possess the requisite capacities for being normal and fully cooperating members of cooperation. Only so will they be able to effectively enjoy the benefits of cooperation. For this reason, Rawls defines two moral powers as the necessary and sufficient condition for being counted a full and equal member of a cooperative society. They are the capacity for a sense of justice and the capacity for a conception of the good. The capacity for a sense of justice is the capacity to understand, to apply, and normally to be moved by an effective desire to act from the principle of justice. The capacity for a conception of the good is the capacity to form, to revise, and rationally to pursue a conception of the good. (PL:302) We are free and equal participants by virtue of having the requisite minimum degree of these two moral powers. We can note that the later Rawls does not ground moral personality on a Kantian interpretation of human nature. Rather, it is a functional idea closely related to the notion of fair social cooperation.[20] We need not appeal to any metaphysical account to justify a conception of moral personality.

Based on the above discussion, we can conclude that rational people have a regulative desire to be effective participants in fair cooperation. This entails that they have a higher-order desire to develop their two moral powers through acting on the principles of justice which can most effectively express the very idea of fair cooperation. Based on this, Rawls can then claim that his principles of justice are better than other alternatives to achieve potential congruence. In the following section, we will use the first principle of liberty as an example to demonstrate this

---

[20] I am indebted to Thomas Scanlon for discussion about this issue.

argument.

## 4 The Good of Basic Liberties

Rawls's first principle of justice states that "each person has an equal right to a fully adequate scheme of equal basic liberties which is compatible with a similar scheme of liberties for all." (PL:291) These liberties include freedom of thought and liberty of conscience, the political liberties and freedom of association, as well as the freedoms specified by the liberty and integrity of the person. Now what we attempt to survey is whether participants have sufficient reasons to view basic liberties as a regulative good for their lives. Below I will present three arguments for the priority of liberty of conscience proposed by Rawls to demonstrate how congruence is possible.[21] All of them are related to the idea of social cooperation.

The first argument holds that liberty of conscience guarantees each participant an equal right to realize his determinate conception of the good. This argument is straightforward and forceful. As I have repeatedly shown, the fact that people have a fundamental interest in pursuing their conception of the good gives rise to the circumstances of justice and constitutes a major reason for social cooperation. A conception of the good consists of a person's final ends and commitments which give substance to his life. Its importance is like what Williams calls a *ground project* or set of projects which are "closely related to his

---

[21] It should be noted that in Rawls's context these arguments are presented to rational parties in the original position of the first stage. However, I believe that they are also valid in the second stage where participants know their conceptions of the good. For Rawls's discussion see "The Basic Liberties and Their Priority," collected in PL:289-371.

existence and which to a significant degree give a meaning to his life."[22] Therefore, no one can afford to sacrifice his conception of the good for other interests. If they were to gamble in this way, Rawls says, "they did not know what a religious, philosophical, or moral conviction was" (PL:311) So, taking conceptions of the good seriously requires us to give liberty of conscience a regulative priority over other desires and preferences in our motivational set.

The second and third arguments are both related to the capacity for a conception of the good. They present the instrumental and intrinsic aspects of this capacity. The second argument holds that the adequate development and exercise of the capacity to form, to revise, and to pursue a conception of the good is an important means to a person's good. The reason is this. As shown in the first argument, we have an essential interest in leading a good life in accordance with our philosophical and religious beliefs. We exercise our rational capacity to form our ends and choose the most effective means to realize them. Without this capacity, we do not know what matters to us and how to pursue our rational good.

Moreover, we also recognize that "leading a good life is different from what we currently believe to be good."[23] For we may be mistaken about the value of what we are doing. We may come to see that our deeply held goals and projects are actually wrong. After deliberation, we may change the final ends of our life. As autonomous and reflective beings, we keep on questioning our values because we worry about whether those values are really worth pursuing. As Rawls puts it,

---

[22] Williams, "Persons, Character, and Morality" in *Moral Luck* (Cambridge: Cambridge University Press, 1981), p.12.

[23] Kymlicka, *Liberalism, Community and Culture* (Oxford: Clarendon Press, 1989), p.10.

"there is no guarantee that all aspects of our present way of life are the most rational for us and not in need of at least minor if not major revision." (PL:313) Because of this, we have a higher-order interest to make sure that there are sufficient freedoms for each of us to questions our beliefs, to examine our values in light of whatever information and arguments available, and to revise our existing projects.

The third argument claims that the effective exercise of the capacity for rational deliberation is not only a means to, but also an essential part of a determinate conception of the good. Living autonomously is itself intrinsically valuable. The main idea comes from Mill's conception of individuality. As autonomous and independent persons, we want to be our own masters and live our lives in our own way. Therefore, in forming and pursuing our conception of the good, we do not want to simply copy it from others. We do not see ourselves as a machine to be built after a model. On the contrary, we are eager to exercise our rational power to make choice. As Mill famously proclaims, "if a person possesses any tolerable amount of common sense and experience, his own mode of laying out his existence is the best, not because it is the best in itself, but because it is his own mode."[24] One of the preconditions of leading a good life is thus that "we lead our life from inside, in accordance with our beliefs about what gives value to life."[25] Rawls fully identifies himself with this liberal tradition by adding that "in addition to our beliefs being true, our actions right, and our ends good, we may also strive to appreciate why our beliefs are true, our actions right, and our ends

---

[24] J. S. Mill, *On Liberty* (New York: Macmillan, 1956), p.82.

[25] Kymlicka, *Liberalism, Community and Culture*, p.13.

good and suitable for us." (PL:312, author's emphasis)

With these arguments to hand, Rawls offers a very powerful defense for the priority of basic liberties. Without basic liberties, we will not be able to fully develop our capacity for a conception of the good. Without effective exercise of this moral power, we will not be able to lead a good life and to be a full participant in cooperation for mutual benefit. Therefore, it is rational for participants to accept the overridingness of the sense of justice. Abiding by the first principle is itself a regulative good for rational beings.

However, we should note that this argument is incompatible with political liberalism. If we commit to viewing personal autonomy as a higher-order value guiding our life, it cannot just be limited to the political domain. It applies to our whole life. That means that if people are convinced by the above arguments, they must have already accepted a liberal conception of the free person. Liberalism takes toleration seriously. But toleration has a limit. The limit is defined by the principle of equal liberties. Different comprehensive doctrines must liberalize themselves and respect other people's liberties in the first place if they want to survive and flourish in a liberal society. A liberal should indeed hold that a liberal conception of the person with its insistence on the priority of equal liberties deserves our allegiance because it is an essential means to, and a constitutive part of, leading a good life. Rawls has no reason to regret this.

## 5 Moral Equality and the Difference Principle

We now turn to examine the possibility of congruence for the difference principle. Predictably, this is much more difficult and controversial than the

argument for the principle of equal liberty. For it requires unequal distribution of income and wealth among participants. There are two fundamentally different attitudes towards this principle. For some, it is too demanding because it sets a very rigid constraint on economic inequalities. But for others, it is not egalitarian enough because it allows unequal distribution among equal participants. This is a complex issue that has aroused heated debate in the past several decades. My discussion will only focus on a specific question: do rational people have an overriding motive to accept the difference principle in a fair cooperation for reciprocity? If they do, congruence may be possible even in this difficult area.

The difference principle stipulates that social and economic inequalities are permissible if and only if it is to the greatest benefit of the least advantaged. Rawls holds that this principle expresses an egalitarian conception of justice. Offhand it is not clear why it is so. For even if a society satisfies the principle, inequalities will still exist and we do not know how large the gap may be. Rawls assures us that it is egalitarian because it embodies a Kantian conception of equality. So before we ask whether rational persons would have enough reason to adopt this principle, we had better work out its moral ground first. It should be noted that I make no attempt to present Rawls's argument from the standpoint of the original position. Rather, I will take the difference principle as directly derived from a conception of moral equality. For, according to Rawls, "to accept the principles that represent a conception of justice is at the same time to accept an ideal of the person; and in acting from these principles we realize such an ideal."[26]

---

[26] Rawls, "A Kantian Conception of Equality" in his *Collected Papers*, ed. Samuel Freeman (Cambridge, Mass: Harvard University Press, 1999), pp.254-55.

The argument for the difference principle can be roughly formulated as follows:

1. All participants in cooperation are equal moral persons.

2. The basis of equality is defined by their having a capacity for a conception of the good and a capacity for a sense of justice to a certain minimum degree. (TJ:505/442 rev.)

3. That they are equal is expressed by the supposition that they each have, and view themselves as having, a right to equal respect and consideration in determining the principles for the basic structure of society.[27]

4. It can be inferred from (3) that justification of a conception of justice should not be troubled by the influence of either social contingencies or natural distribution of abilities and talents. These factors are arbitrary from a moral point of view, and will affect the status of equal moral personality in determining the principles of justice.[28] (TJ:74-75/64-65 rev.)

5. Since all participants are equal, they are granted a veto. Any

---

[27] Rawls, "A Kantian Conception of Equality," p.255.

[28] I am aware that there are other interpretations of "arbitrary from the moral point of view." For example, Rawls sometimes says that it is arbitrary because no one deserves his place in the distribution of natural endowments and his initial starting position in society. (TJ:104/87 rev.) Or as luck egalitarians claim, it is arbitrary because they are not the outcome of people's choice. No one should be responsible for this kind of brute luck. I believe that these three interpretations are conceptually different from one another, and will result in a very different understanding of the difference principle. In my view, the interpretation of moral equality is the strongest one. For if we drop equality, the other two interpretations alone are not enough to support the claim that each person should be treated as equal. Desert and luck/choice dichotomy are conceptually compatible with unequal treatment. Moreover, Rawls clearly remarks that "once we try to find a rendering of them which treats everyone equally as a moral person, and which does not weight men's share in the benefits and burdens of social cooperation according to their social fortune or their luck in the natural lottery, it is clear that the democratic interpretation is the best choice among the four alternatives." (TJ:75/65 rev.) The term "luck egalitarianism" is drawn from Elizabeth S. Anderson, "What is the Point of Equality?" *Ethics* 109 (1999), pp. 287-337. Luck egalitarians include Ronald Dworkin, *Sovereign Virtue: The Theory and Practice of Equality* (Cambridge, Mass.: Harvard University Press, 2000); Thomas Nagel, *Equality and Partiality* (New York: Oxford University Press, 1991); Eric Rakowski, *Equal Justice* (Oxford: Clarendon Press, 1991); Richard J. Arneson, "Luck Egalitarianism and Prioritarianism," *Ethics* 110 (2000), pp.339-349.

principles of justice must therefore be justified to every rational member of the cooperative scheme. "In this sense, it is egalitarian." (TJ:103)[29]

6. Following (5), the only acceptable principle seems to be that every participant should have an equal share of income and wealth.

7. However, it is irrational to stop at equal division if there is an alternative scheme that can make all participants better off, including the expectations of the least advantaged.

8. The least advantaged recognize that if they do not give the better off higher economic benefits, their long-term prospects will not be improved. The more attractive prospect of those better off "acts as incentives so that the economic process is more efficient, innovation proceeds at a faster pace, and so on." (TJ:78/68 rev.) Or put it in a third-person perspective, "society must take organisational requirements and economic efficiency into account."[30]

Comment [LC1]: Ibid?

9. Thus we arrive at the difference principle. "Taking equality as the basis of comparison those who have gained more must do so on terms that are justifiable to those who have gained the least."[31]

We can see that the whole argument begins with a conception of moral equality. The difference principle is not presented as a result of rational choice behind the veil of ignorance. Participants are supposed to know their relative social and natural advantages and their conceptions of the good; they accept the difference principle because they are convinced by its moral premises. Rawls

---

[29] This sentence is deleted in the revised edition.

[30] Rawls, "A Kantian Conception of Equality," p.262.

[31] Rawls, "A Kantian Conception of Equality," p.262.

sometimes suggests that this argument only expresses some fundamental intuitive ideas informally because in a contract theory, "all arguments, strictly speaking, are to be made in terms of what it would be rational to agree to in the original position." (TJ:75/65 rev.) I do not agree with this account. Conversely, I find this argument more fundamental and effective than the rational choice argument. It is more fundamental because the difference principle is directly derived from a conception of moral equality, according to which the contractual constraints of the original position are defined. In this sense, the contractarian argument is redundant. As Charvet rightly points out, "if this equality of value and rights is a constraining condition on the choice situation, which has to be independently justified, then it would be pointless to present the argument for justice in contractarian form."[32] The role of contract is merely to serve as a device of representation to work out the implication of moral equality which has been antecedently justified. It is more effective because we need not bother with the plausibility of the maximin rule and the disputable motivational assumption of mutual disinterest. Since the argument straightforwardly appeals to our moral beliefs, its egalitarian character and moral attractiveness are more readily observed than those of the rational choice argument.

This argument has its own motivational assumptions though. It presumes that all participants have a desire to accept equality as the most fundamental value in determining the terms of cooperation. Otherwise, they would have opted for a conception of justice as mutual advantage in accordance with their relative

---

[32] John Charvet, *The Idea of an Ethical Community* (Ithaca & London: Cornell University Press, 1995), p.167.

bargaining power. The question concerned is how a participant's life will be shaped if he takes up the impartial standpoint in practical reasoning for the fundamental principle of justice. Following the claim of the unity of practical reasoning, the respect for equality must occupy a regulative place in his conception of the good. The acceptance of moral equality presupposes a corresponding egalitarian motive. Moreover, participants understand and accept the implication of social cooperation, namely they are willing to give priority to the terms of cooperation provided that others will act in the same way. Thus, for a rational agent to be convinced of Rawls's argument, he must identify himself as a full egalitarian. Without this assumption, justification of the difference principle would not even be initiated.

Be that as it may, it would follow from this line of reasoning that moral equality does not deductively lead to the difference principle. It actually requires an additional psychological assumption that the better off need extra economic incentives to work more efficiently for the common good of the whole community. Not allowing inequality in expectation makes the least advantaged even worse off. For otherwise, the better off would lack motive to develop new technology and increase production. Participants finally opt for the difference principle because they take this incentive argument into account. As Rawls remarks, "something of this kind must be argued if these inequalities are to be just by the difference principle." (TJ:78/68 rev.) In other words, without the incentive assumption, participants would stick to the principle of equal share.[33]

---

[33] Cohen has offered an incisive discussion on this issue. "Incentives, Inequality, and Community" in *Equal Freedom* ed. Stephen Darwall (Ann Arbor: The University of Michigan Press), pp.331-397. Also see Cohen, *If You're an Egalitarian, How Come You're So Rich?* (Cambridge,

Nevertheless, the incentive argument is inconsistent with the moral commitment to equality. As shown above, participants are presumed to be full egalitarians accepting that all human lives are of equal value. That is why they accept that no one should have the right to gain from arbitrary natural and social advantages without having others' consent. Since they are moved by this impartial concern, their sense of justice is supposed to be able to outweigh the demand for economic incentives. They are willing to share one another's fate. As full egalitarians, therefore, they should favour a high degree of equal distribution. They do not do so exactly because they qualify their moral commitment in the light of the economic incentive argument.[34] The difference principle is justified on the basis of an admission of the psychological limits of human nature. It is not the most ideal principle of justice. It is a second best that we cannot but accept because of the necessity for economic incentives. Surprisingly, Rawls seems to admit this consequence:

> One might think that *ideally* individuals should want to serve one another. But since the parties are assumed not to take an interest in one another's interests, their acceptance of these inequalities is only the acceptance of the relations in which men stand in the circumstances of justice. They have no grounds for complaining of one another's motives. *A person in the original position would, therefore, concede the justice of these inequalities.* (TJ:151/131 rev., my emphasis)[35]

This paragraph suggests that if rational parties have sufficient moral motive

---

Mass: Harvard University Press, 2000), chap.8.

[34] Rawls mentions that inequality is sometimes allowed to cover the costs of training and education. This account is different from the incentive argument. Strictly speaking, this kind of inequality is still in the framework of equal share because its function is to compensate costs incurred.

[35] The last sentence is deleted from the revised edition.

to help one another, they would not accept a distributive scheme that allows economic inequalities. But since this assumption is too idealistic under the circumstances of justice, people cannot but accept inequalities specified by the difference principle. The principle is therefore a result of compromise rather than the most justifiable political ideal. Needless to say, this conclusion will devastate Rawls's whole philosophical enterprise which aims to justify a most reasonable conception of justice grounded on freedom and equality.

Rawls may defend his position by saying that the assumption of mutual disinterest and the need for economic incentive only apply to the original position. This reply misses the point. When the parties make their rational choice behind the veil of ignorance, they know that the principles will apply to the basic structure of society. They also know the laws of human psychology. More importantly, they know that they will act on an effective sense of justice in a well-ordered society. "Once the veil of ignorance is removed, the parties find that they have ties of sentiment and affection, and want to advance the interests of others to see their ends attained." (TJ:129/111 rev.) So in making their decision, rational parties have no need to concede the justice of inequalities in order to "set up various incentives which succeed in eliciting more productive efforts." (TJ:151)[36] Furthermore, this response is inapplicable to my argument because it does not take place in the original position. Participants are presumed to be moved by a commitment to moral equality.

Alternatively, Rawls may bite the bullet and argue that the claim of economic

---

[36] This sentence is deleted from the revised edition.

incentive is a psychological fact that we should take for granted in constructing a conception of justice. We are inevitably subject to the conditions of human life with moderate scarcity and competing claims. We are not living in a Kantian kingdom of ends in which each rational agent is expected to have the purest sense of duty to act in accordance with the categorical imperative. On the contrary, "justice as fairness is a theory of human justice and among its premises are the elementary facts about persons and their place in nature." (TJ:257/226 rev.) The ideal of moral equality must be subject to the limitations of the circumstances of justice, and the difference principle is a compromise that we should bear with. I am still not convinced by this defence. It is incompatible with the egalitarian character of justice as fairness. Rawls can no longer claim that justice is the first virtue of social institutions because the difference principle fails to express a most justifiable conception of justice from a moral point of view. I believe that Rawls is unsatisfied with this argument also. Otherwise, he would not delete several citations concerning the economic incentive mentioned above from the revised edition of *A Theory of Justice.*

Is there any alternative argument to justify the difference principle in the egalitarian framework without appealing to the incentive argument then? Obviously, if there is one, it must be a moral argument compatible with the equal worth of human beings. The difference principle can then be presented not as a compromise, but as a result of moral agreement. Nagel's idea of "self-limiting modifications of impersonal morality" shows promise in this regard.[37]

---

[37] Nagel, *The View from Nowhere*, p.204. Nagel further develops his argument in *Equality and Partiality* (New York: Oxford University Press, 1991).

According to Nagel, we can think of impersonal morality as developing in stages. At the first stage, we accept moral equality and recognize that objectively we are no more important than anyone else. So in practical reasoning our interests and welfare are accorded as much weight as those of other people. This impartial standpoint requires that each person has a right to equal respect and consideration in determining the principles of justice. Apparently, this is exactly the starting point of justice as fairness which justifies equal distribution of primary social goods.

With further reflection on human motives, we are urged to go one step further to modify the principle of equal share. First of all, even viewing the situation from an impartial perspective, participants of cooperation may recognize that as autonomous and independent agents, they have fundamental interests in forming and pursuing their conceptions of the good which contain fundamental human needs as well as the major activities, projects and commitments around which their lives are organized. These interests provide meaning to their lives and set the background for their practical reasoning. Thus, they accept that it is morally legitimate for an individual, within certain impartial limits, to devote disproportionate attention to those things that matter most to him. The very fact of their importance to his life provides by itself a reason for justification. So we are not morally required to evaluate and decide our actions from a strictly first-order impartial point of view. This is, however, not a compromise between equality and self-interest. Rather, it is a judgment made within the framework of impartiality. "When we regard people objectively and think about how they should live, their

motivational complexity is a consideration."[38] Expecting people to sacrifice their ground projects and attachments to their loved ones for the sake of impartial considerations *per se* is unreasonable and excessively demanding. As Scheffler suggests, a moderate conception of morality should strive to strike a balance between the following two propositions:

> The first proposition is that, from an impersonal standpoint, everyone's life is of equal intrinsic value and everyone's interests are of equal intrinsic importance. The second proposition is that each person's interests nevertheless have a significance for him or her that is out of proportion to their importance from an impersonal standpoint.[39]

Now suppose that rational participants accept the modification of impersonal morality and take the personal standpoint into account in deciding principles of justice. They then have a moral argument for a certain degree of unequal distribution of income which is not a concession to the limits of human weakness. They recognize that income and wealth are important all-purpose means to realize their conception of the good to which they may legitimately devote disproportionate attention. Economic inequalities are thus, within limits, morally permissible. Those better off have a legitimate expectation of more reward for exercising their natural talents. This does not mean that the dispensation for inequalities is unlimited. For the principles of justice should also be sensitive to the guiding principle that all participants are of equal value regardless of the

---

[38] Nagel, *The View from Nowhere*, p.202. But Nagel himself seems to view the modification as a compromise based on "tolerance and recognition of limits" between "our higher and lower selves in arriving at an acceptable morality." I believe that this account is inconsistent with Nagel's overall view about the co-existence of objective and subjective standpoints. A similar critique can also be found in Scheffler, *Human Morality*, p. 125.

[39] Scheffler, *Human Morality*, p.122

unequal distribution of natural talents and social background. The difference principle represents a balance between the impersonal value of others and our naturally disproportionate concern for our own lives.

The modification of impartiality can somewhat alleviate the tension between the personal and impersonal point of view. Compared to the principle of equal share, the difference principle is less demanding and more motivationally accessible to normal rational persons. For participants who have already accepted society as a fair system of cooperation for mutual advantage between free and equal persons, this is an arrangement that they can reasonably accept. Therefore, congruence between the difference principle and participants' conceptions of the good is a realistic possibility. That being said, a Rawlsian well-ordered society is still an ideal very far from our existing society. Rawls's theory demands an economic distribution deriving from a strong commitment to moral equality, which is in strong tension with a capitalist market economy. How to cultivate an egalitarian ethos through moral education and reform of the basic structure is a prerequisite for a liberal egalitarian society, and for the moral stability of justice as fairness.[40]

---

[40] For this issue, see Cohen, *If You're an Egalitarian, How Come You're So Rich?* (Cambridge, Mass: Harvard University Press, 2000), chap.10. Scheffler also provides a good reflection on it. See *Human Morality*, chap.8.

# CONCLUSION

In the introduction to this thesis, I describe my project as an attempt to vindicate the importance of moral stability in liberal justification. I hope that my arguments have affirmed this claim. I will now summarize some main points that I believe this research has contributed to the study of Rawls.

I have made it clear at the outset that the problem of stability is essentially concerned with the motivational priority of a sense of justice. It is about how a rational agent, having a fundamental interest in advancing his conception of the good, could have sufficient motive to act morally. Rawls recognizes that justice and goodness represent two distinct standpoints in practical reasoning. Both of them make claims on us. However, a tension between these two standpoints is inherent in Rawls's theory. For the subjective circumstances of justice presume that participants of cooperation have different conceptions of the good; it is a natural and legitimate desire for them to pursue their good and in so doing, they make conflicting claims on the distribution of natural and social resources. If Rawls wants to justify the overridingness of justice, he must take up the problem of stability and demonstrate why it is rational for an agent to give precedence to moral considerations over his ends and goals.

This is what Korsgaard calls the normative question of moral philosophy. "The normative question is a first-person question that arises for the moral agent who must actually do what morality says."[1] Thus, stability is a justificatory

---

[1] Christine Korsgaard, *The Sources of Normativity*, ed. Onora O'Neill (Cambridge: Cambridge University Press, 1996), p.16.

problem that determines the desirability of justice as fairness. It is wrong for Rawls and most critics to say that the goal of stability is to affirm the feasibility of a conception of justice independently justified.[2] This account cannot make sense of Rawls's own claim that stability is essential to justifying justice as fairness. Worse still, it will lead Rawls to commit a category mistake.[3] That is why I stress in Chapter One that we must draw a clear distinction between social stability and moral stability. Moral stability is not concerned with a purely practical matter of social order which is of no relevance to the justifiability of a conception of justice. Once this distinction is established, most criticism stemming from this misunderstanding can be dismissed. This is the first major claim I have made in this thesis.

My second major claim is about the place of stability in Rawls's two-stage justificatory structure. Against Rawls's own account, I have argued that stability is the concern of both stages. It is one of the main grounds for contractors in the original position to prefer Rawls's two principles to the principle of average utility. It is therefore misleading for Rawls to say that the problem of stability arises in the second stage only after the principles of justice have been worked out in the first one on independent grounds. In addition, I contend that the real force moving the contractors to adopt the maximin rule actually results from moral considerations. Rawls's principles are the result of a moral argument rather than

---

[2] For example, Freeman, perhaps the most sympathetic critic on this issue, says that "the question of its stability is raised to test the *feasibility* of a just society conceived along the lines of this conception." "Congruence and the Good of Justice" in *The Cambridge Companion to Rawls* ed. Samuel Freeman (Cambridge: Cambridge University Press, 2003), p.279, my emphasis.

[3] I must say that Rawls is partly responsible for this consequence because he fails to distinguish these two conceptions of stability when the problem was first formulated in *A Theory of Justice*. His later philosophical development, however, indicates that he is well aware of this distinction. See PL:142

rational calculation. Justice as fairness, strictly speaking, is not a contractarian theory.

Furthermore, I have explained that the second stage is necessary because the priority of the sense of justice can only be settled when people are allowed to have full knowledge of their conceptions of the good. As Rawls remarks, "the problem is whether the regulative desire to adopt the standpoint of justice belongs to a person's own good when viewed in the light of the thin theory with no restriction on information." (TJ:567/497 rev.) This implies that the justification of justice as fairness is unfinished in the first stage. It needs one more step. The principles derived from the original position must be shown to be stable in the second stage. My argument is confirmed by Rawls's later remark that "the argument for the principles of justice is not complete until the principles selected in the first part are shown in the second part to be sufficiently stable." (PL:141) In that case, the decision in the original position is no longer the final court of justification.

My third major claim is about Rawls's congruence argument. Any discussion on the overridingness of moral motive must involve an account of practical reasoning. I point out that Rawls has adopted a desire-based prudential rationality and the idea of a rational plan of life to explain our reasons for action. When this view is combined with Rawls's internalist position, it naturally leads to a congruence argument. Congruence is close to what Korsgaard calls the idea of "reflective endorsement." This view holds that morality is grounded in human nature. When an explanation of human nature is found, it can then be argued that those moral principles that best express our nature are good for us. We would therefore have sufficient motive to accept the claim of morality. The priority of the

sense of justice is grounded on the harmony of two normative points of view, morality and self-interest.[4] Rawls's appeal to the Kantian interpretation of human nature as free and equal rational being as the basis of congruence is such a model.

I have argued that this metaphysical interpretation of human nature has turned Rawls into a liberal perfectionist within a classical teleological framework, rendering Rawls's dichotomy between teleology and deontology misleading. If the moral foundation of justice as fairness is a Kantian conception of the person, it is hard for Rawls to claim that he upholds a position of liberal neutrality. Moreover, this interpretation is inconsistent with Rawls's desire-based conception of prudential rationality. This explains why the later Rawls is forced to make a philosophical shift to political liberalism.

Finally, my last major claim holds that the approach of potential congruence is more desirable and feasible than the idea of an overlapping consensus for affirming the motivational priority of the sense of justice. Given the notion of the unity of practical reasoning and the pervasiveness of moral feelings, congruence is a realistic project for justice as fairness even though there is no room for it to be taken for granted conceptually in a post-metaphysical era. The later Rawls has given up this hope because of the fact of reasonable pluralism. However, I do not see the prospect of constructing a freestanding and thin conception of liberal egalitarianism that calls for the whole-hearted allegiance of citizens who hold a

---

[4] Korsgaard, *The Sources of Normativity*, pp.19, 60. It is interesting to note that Korsgaard ascribes Rawls's congruence argument to the view of reflective endorsement while holding that Rawls's Kantian constructivism belongs to the view of "the appeal to autonomy." Korsgaard does not explain how these competing views can coexist in Rawls's account of normativity. A plausible explanation is that she does not view congruence as an essential argument for justice as fairness. However, my analysis shows that Rawls's ultimate answer to the normative question rests on his argument for the reflective endorsement.

diversity of liberal and non-liberal comprehensive doctrines. I believe that the motivational priority of justice can only be grounded on a substantive moral ideal. Toleration and mutual respect are better justified by appealing to a liberal ideal of autonomy and equality. It is undoubtedly comprehensive. But liberalism should enable citizens to lead a liberal way of life. Only when citizens share the liberal ideal and recognize the fundamental good of living an ethical life can a Rawlsian well-ordered society claim to be realistically utopian.[5] Given that we have a moral nature, we have no reason to give up that hope.

---

[5] The idea of political philosophy as realistically utopian is first raised in Rawls, *Justice as Fairness* ed. Erin Kelly (Cambridge, Mass: Harvard University Press, 2001), p.4.

# BIBLIOGRAPHY


Arthur, J. & Shaw, W. (1991). (ed.) *Justice and Economic Distribution* (Englewood Cliffs, New Jersey: Prentice Hall).

Ackerman, B. (1980). *Social Justice in the Liberal State* (New Haven: Yale University Press).

Alejandro, R. (1993). "Rawls's Communitarianism," *Canadian Journal of Philosophy* 23, pp. 75-100.

Anderson, E. (1999) "What is the Point of Equality?" *Ethics* 109, pp. 287-337.

Arneson, R. (1982). "The Principle of Fairness and Free-Rider Problems," *Ethics* 92, pp.616-33.

——(1990). "Neutrality and Utility", *Canadian Journal of Philosophy* 20, pp. 215-240.

—(1996). "Responsibility, Neutrality, and Political Liberalism", typescript.

—(2000). "Luck Egalitarianism and Prioritarianism", *Ethics* 110, pp. 339-349.

Avineri, S. & de-Shalit, A. (1992). (ed.) *Communitarianism and Individualism* (New York: Oxford University Press).

Baier, K. (1989). "Justice and the Aims of Political Philosophy", *Ethics* 99, pp. 771-790.

Barry, B. (1973). *The Liberal Theory of Justice* (Oxford: Clarendon Press).

—(1989). *Theories of Justice* (California: University of California Press).

—(1990). *Political Argument: A Reissue with a New Introduction* (Berkeley and Los Angeles: University of California Press).

—(1991). *Liberty and Justice: Essays in Political Theory* (Oxford: Clarendon Press.).

—(1995a). *Justice as Impartiality* (Oxford: Clarendon Press).

—(1995b). "John Rawls and the Search for Stability", *Ethics* 105, pp. 874-915.

——(1998). "Something in the Disputation not Unpleasant " in *Impartiality, Neutrality and Justice* ed. P. Kelly (Edinburgh: Edinburgh University Press), pp. 186-257.

Bates, S. (1974). "The Motivation to be Just", *Ethics* 85, pp. 1-17.

Berlin, I. (1969). *Four Essays on Liberty* (Oxford: Clarendon Press).

—(1991). *The Crooked Timber of Humanity* (London: Fontana Press).

Buchanan, A. (1989). "Assessing the Communitarian Critique of Liberalism" *Ethics* 99, pp. 852-882.

Caney, S. (1995). "Anti-perfectionism and Rawlsian Liberalism", *Political Studies* 43, pp. 248-264.

Charvet, J. (1981). *A Critique of Freedom and Equality* (Cambridge: Cambridge University Press, 1995).

—(1995). *The Idea of an Ethical Community* (Ithaca: Cornell University Press).

Cohen, G.A. (1989). "On the Currency of Justice", *Ethics* 99, pp.906-944.

—(1995) "Incentives, Inequality, and Community" in *Equal Freedom* ed. S. Darwall (Ann Arbor: The University of Michigan Press), pp.331-397.

—"Rescuing Justice from Constructivism," typescript.

—(2000) *If You're an Egalitarian, How Come You're So Rich?* (Cambridge, Mass: Harvard University Press).

——(2003). "Facts and Principles," *Philosophy and Public Affairs* 31, no.3, pp.211-45.

Cohen, J. (1989). "Democratic Equality", *Ethics* 99, pp.727-51.

—(1993). "Moral Pluralism and Political Consensus" in *The Idea of Democracy* ed. D. Copp, J. Hampton and J. Roemer (Cambridge: Cambridge University Press), pp. 270-91.

Copp, D. (1996) "Pluralism and Stability in Liberal Theory", *The Journal of Political Philosophy* 4, pp.191-206.

Corlett, J.A. (1991). (ed.). *Equality and Liberty: Analyzing Rawls and Nozick* (London: Macmillan).

Daniels, N. (1975). (ed.). *Reading Rawls* (Stanford, California: Stanford University Press).

Davion, V. & Wolf, C. (2000). (ed.). *The Idea of a Political Liberalism: Essays on Rawls* (Lanham, Maryland: Rowman & Littlefield).

Darwall, S. (1997). "Reasons, Motives, and the Demands of Morality: an Introduction", in *Moral Discourse and Practice: Some Philosophical Approaches* ed. S. Darwall, A. Gibbard, P. Railton (New York: Oxford University Press), pp.305-12.

Delue, S. (1980). "Aristotle, Kant and Rawls on Moral Motivation in a Just Society", *The American Political Science Review* 74, pp.385-393.

Doppelt, G. (1989). "Is Rawls's Kantian Liberalism Coherent and Defensible", *Ethics* 99, pp. 815-851.

Dworkin, R. (1975) "The Original Position" in *Reading Rawls*, ed. N. Daniels (Stanford, California: Stanford University Press), pp. 16-53.

—(1977). *Taking Right Seriously* (London: Duckworth).

——(1978). "Liberalism" in *Public and Private Morality*, ed. S. Hampshire (Cambridge: Cambridge University Press), pp. 113-43.

—(1985). *A Matter of Principle* (Oxford: Clarendon Press).

—(1990). "Foundations of Liberal Equality" in *The Tanner Lecture on Human Values Vol. XI* (Salt Lake City: University of Utah Press), pp.3-119.

—(2000). *Sovereign Virtue: The Theory and Practice of Equality* (Cambridge, Mass.: Harvard University Press).

Engstrom. S and Whiting, J. (1996). (eds.) *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty* (New York: Cambridge University Press).

Estlund, D. (1996). "The Survival of Egalitarian Justice in John Rawls's *Political Philosophy* ", *The Journal of Political Philosophy* 4, pp. 68-78.

Flanagan, O. (1991). *Varieties of Moral Personality* (Cambridge, Mass: Harvard University Press).

Foot, P. (1981). *Virtues and Vices* (Oxford: Basil Blackwell).

Frankena, W. (1973) *Ethics* (Englewood Cliffs, New Jersey: Prentice Hall).

Frankfurt, H. (1988). *The Importance of What We Care About* (Cambridge: Cambridge University Press).

Freeman, S. (1991). "Contractualism, Moral Motivation, and Practical Reason", *The Journal of Philosophy* 88, pp. 281-303.

——(1994). "Political Liberalism and the Possibility of a Just Democratic Constitution", *Chicago-Kent Law Review* 69, pp. 619-68.

—(2003a). (ed.). *The Cambridge Companion to Rawls* (Cambridge: Cambridge University Press).

—(2003b). "Congruence and the Good of Justice" in *The Cambridge Companion to Rawls*, pp. 277-315.

Galston, W. (1989). "Pluralism and Social Unity", *Ethics* 99, pp. 711-726.

—(1991). *Liberal Purposes* (Cambridge University Press).

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

Gibbard, A. (1991). "Constructing Justice", *Philosophy and Public Affairs* 20, pp. 264-279.

Gray, J. (2000a). *Two Faces of Liberalism* (Cambridge: Polity Press).

Gutmann, A. (1985). "Communitarian Critics of Liberallism", *Philosophy and Public Affairs* 14, pp. 308-322.

Habermas, H. (1995). "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism", *The Journal of Philosophy*, 92, pp. 109-31.

Hampton, J. (1980). "Contracts and Choice: Does Rawls Have a Social Contract Theory?" *The Journal of Philosophy* 77, pp.315-38.

—(1989). "Should Political Liberalism Be Done without Metaphysics?" *Ethics* 99, pp.791-814.

—(1993a). "The Moral Commitments of Liberalism" in D. Copp, J. Hampton, and J. Roemer (eds.) *The Idea of Democracy* (Cambridge: Cambridge University Press), pp.292-313.

—(1993b). "Contract and Consent" in R.Goodin & P. Pettit (eds.) *A Companion to Political Philosophy* (Oxford: Blackwell), pp. 379-393.

—(1994). "The Common Faith of Liberalism", *Pacific Philosophical Quarterly* 75 (1994), pp.186-216.

Harsanyi, J. (1975). "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory", American Political Science Review 69, pp.594-606; collected in *John Rawls: Critical Assessment of Leading*

*Political Philosophers*, vol.1, pp.216-38.

Hart, H. L. A. (1975) "Rawls on Liberty and its Priority" in *Reading Rawls* ed. Norman Daniels (Stanford, California: Stanford University Press), pp. 230-52.

Herman, B. (1993). *The Practice of Moral Judgment* (Cambridge, Mass: Harvard University Press).

Hill, T. (1994). "The Stability Problem in *Political Liberalism*", *Pacific Philosophical Quarterly* 75, pp. 333-352.

Hobbes, T. (1991). *Leviathan* ed. R. Tuck (Cambridge: Cambridge University Press).

Horton, J. (1996). "Toleration as a Virtue" in *Toleration: An Elusive Virtue* ed. D. Heyd (Princeton, New Jersey: Princeton University Press).

Hume, D. (1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L.A.Selby-Bigge (Oxford: Oxford University Press).

Hurka, T. (1998) "Perfectionism" in *Routledge Encyclopedia of Philosophy,* ed. E. Craig,(London: Routledge, 1998), retrieved from http://www.rep.routledge.com/article/L070SECT5.

Johnson, D. (1994). *The Idea of a Liberal Theory* (Princeton, New Jersey: Princeton University Press).

Kant, I. (1996). *Groundwork of the Metaphysics of Morals* collected in *Practical Philosophy*, tran. & edited by M. Gregor (Cambridge: Cambridge University Press).

——(1970). *Political Writings*, trans. H. B. Nisbet (Cambridge: Cambridge University Press).

Kelly, P. (1998). (ed.). *Impartiality, Neutrality and Justice* (Edinburgh: Edinburgh University Press).

Klosko, G. (1994). "Rawls's Argument from Political Stability", *Columbia Law Review* 94, pp.1882-1897.

Korsgaard, C. (1996a). *The Sources of Normativity* ed. Onora O'Neill (Cambridge: Cambridge University Press).

—(1996b). *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press).

—(1998) "Teleological Ethics," in *Routledge Encyclopedia of Philosophy* ed. E. Craig (London: Routledge, 1998), retrieved from http://www.rep.routledge.com/article/L103.

Krasnoff, L. (1998). "Consensus, Stability, and Normativity in Rawls's Political Liberalism", *The Journal of Philosophy* 95, pp. 269-292.

Kukathas, C. & Pettit, P. (1990). *Rawls: A Theory of Justice and its Critics* (Cambridge: Polity Press).

Kukathas, C. (2003). (ed.) *John Rawls: Critical Assessment of Leading Political Philosophers* 4 vols. (London & New York: Routledge).

Kymlicka, W. (1989). *Liberalism, Community and Culture* (Oxford: Clarendon Press).

—(1990). *Contemporary Political Philosophy* (New York: Oxford University Press).

—(1992). "Liberal Individualism and Liberal Neutrality" in *Communitarianism and Individualism*, ed. S. Avineri and De-Shali.A (Oxford University Press).

—(1995a). (ed.) *The Rights of Minority Cultures* (Oxford: Oxford University Press).

—(1995b). *Multicultural Citizenship* (Oxford: Clarendon Press).

Larmore, C. (1987). *Patterns of Moral Complexity* (Cambridge: Cambridge University Press).

—(1996). *The Morals of Modernity* (Cambridge: Cambridge University Press).

—(1999) "The Idea of a Life Plan" in *Human Flourishing* eds. E. Paul, F. Miller & J. Paul (Cambridge: Cambridge University Press), pp.96-112.

MacIntyre, A. (1981). *After Virtue* (London: Duckworth).

Marneffe, P. (1990). "Liberalism, Liberty, and Neutrality", *Philosophy and Public Affairs* 19, pp. 251-274.

Martin, R. (1985). *Rawls and Rights* (Lawrence: University Press of Kansas).

Matravers, M. (2000). *Justice and Punishment* (Oxford: Oxford University Press).

McClennen, E. (1989). "Justice and the Problem of Stability", *Philosophy and Public Affairs* 18, pp.3-30.

Mendus, S. (1989). *Toleration and the Limits of Liberalism* (London: Macmillan).

——(1996). "Tragedy, Moral Conflict, and Liberalism" in D. Archard (ed.) *Philosophy and Pluralism* (Cambridge University Press).

—(1999) "The Importance of Love in Rawls's Theory of Justice", *British Journal of Political Science* 29, pp. 57-75.

—(2002). *Impartiality in Moral and Political Philosophy* (New York: Oxford University Press).

Mill, J. S. (1956). *On Liberty* (New York: Macmillan).

Muirhead, J.H. (1932) *Rule and End in Morals* (Oxford: Oxford University Press).

Mulhall, S. and Swift, A. (1992). *Liberals and Communitarians* (Oxford: Blackwell).

Nagel, T. (1970). *The Possibility of Altruism* (Princeton: Princeton University Press).

—(1979). *Mortal Questions* (Cambridge: Cambridge University Press).

—(1986). *The View from Nowhere* (New York: Oxford University Press).

——(1987). "Moral Conflict and Political Legitimacy", *Philosophy and Public Affairs* 16, pp. 215-240.

—(1991). *Equality and Partiality* (Oxford University Press).

Nielsen, K. (1985). *Equality and Liberty* (Totowa: Rowman & Allenheld).