Why Political Liberalism?

On John Rawls's Political Turn

Paul Weithman

Why Political Liberalism?

OXFORD POLITICAL PHILOSOPHY

GENERAL EDITOR: SAMUEL FREEMAN, UNIVERSITY OF PENNSYLVANIA

Oxford Political Philosophy publishes books on theoretical and applied political philosophy within the Anglo-American tradition. The series welcomes submissions on social, political, and global justice, individual rights, democracy, liberalism, socialism, and constitutionalism.

N. Scott Arnold
Imposing Values: An Essay on Liberalism and Regulation

Peter de Marneffe
Liberalism and Prostitution

Debra Satz
Why Some Things Should Not Be for Sale
The Moral Limits of Markets

William J. Talbott
Human Rights and Human Well-being

Paul Weithman
Why Political Liberalism? On John Rawls's Political Turn

Why Political Liberalism?

On John Rawls's Political Turn

Paul Weithman

OXFORD
UNIVERSITY PRESS
2010

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Weithman, Paul J., 1959-
Why political liberalism? : on John Rawls's political turn / by Paul Weithman.
p. cm.
ISBN 978-0-19-539303-3 (alk. paper)
1. Justice. 2. Liberalism. 3. Political stability. 4. Rawls, John, 1921-2002—Criticism and interpretation. I. Title.
JC578.W42 2010
320.092—dc22
2009047179
ISBN: 9780195393033

1 3 5 7 9 8 6 4 2
Printed in the United States of America
on acid-free paper

For my teachers

"When fully articulated, any conception of justice expresses a conception of the person, of relations between persons and of the general structure and ends of social cooperation. To accept the principles that represent a conception is at the same time to accept an ideal of the person, and in acting from these principles, we realize such an ideal."

—John Rawls, "A Kantian Conception of Equality"

Contents

Acknowledgments

This book has been a long time in the making and I have incurred many debts in the course of writing it. I first wrote up part of the book as a contribution to a conference organized at Trinity College, Dublin, in 2006 by Nigel Biggar. I received especially helpful comments at that conference from Nigel himself and from Nicholas Wolterstorff. For many years, both before and after the conference, I have conversed and corresponded with a number of people about Rawls's work. I have benefited greatly from my exchanges with Robert Adams, Robert Audi, Daniel Brudney, Peter de Marneffe, Neil Delaney, Thomas Pogge, Henry Richardson, Fred Rush, James Sterba, and Peter Wicks. In the spring of 2009, an audience at the University of Toronto heard a paper that distilled the central argument of the book. I am grateful to many of those who attended my talk, particularly Ronald Beiner and Melissa Williams, for helpful comments on that occasion. Alasdair MacIntyre read the manuscript of this book in its entirety and offered comments that were exceptionally trenchant and insightful. I owe a special debt to John Roos of the Notre Dame Political Science Department, with whom I have team-taught *A Theory of Justice* for many years, and to the many students who have passed through our course in Notre Dame's Philosophy, Politics and Economics Program. My understanding of Rawls's work would be much less precise if I had not had the privilege of working with John to explain it to those first-rate undergraduates.

It has been an honor and a pleasure to work with Oxford University Press. Comments by Colin Bird and a second, anonymous, reader did much to improve the book; indeed, one of Colin's comments led me to recast large parts of the argument in a form that I believe is considerably more perspicuous. Peter Ohlin of Oxford has been a patient and understanding editor. I am

very fortunate that Samuel Freeman inaugurated his series with Oxford just before I approached the Press about publishing this book and that Peter directed my manuscript to him. Samuel's comments at the final stages did much to improve the book. The book could not have reached publishable form without the patient labors of Natalie Johnson of Oxford and of Alex Jech. I am profoundly thankful for their help.

The long quotation at the beginning of Chapter VII is reprinted by permission of the publisher from *A Theory of Justice* by John Rawls, p. 574-75/503, Cambridge, MA: The Belknap Press of Harvard University Press, Copyright © 1971, 1999 by the President and Fellows of Harvard College.

Much of this book was drafted during an administrative leave I enjoyed after six years as chair of the Notre Dame Philosophy Department. My years of administrative work were, if not all-consuming, at least voracious consumers of my professional life. I am grateful to my colleagues for making those years such rewarding ones, and to the University for funding the leave that followed.

Sustained concentration on a book can impose an enormous strain on one's family. I am grateful to my late mother Patricia Weithman, to my beloved wife Maura Ryan, and to our wonderful children Annie and Meggie for bearing up under the strain without complaint, for overlooking the many things that I left undone while I was writing it, but most of all for their unfailing love.

When my colleague David Solomon first heard about this project, he described it as an exercise in filial piety. He meant that the book is a filial tribute by a Rawls student to his *doktorvater*, and in that he was right. But in describing the book as he did, David spoke more truly than he knew. My interest in philosophy was kindled and sustained by a succession of extraordinary teachers. Rawls and Judith Shklar, who codirected my dissertation, were among them and my debts to the two of them are incalculable, but many others have helped along the way. They include Karl Ameriks, Roderick Firth, Richard Foley, Warren Goldfarb, John Jenkins (now Fr. John Jenkins, CSC), Vaughn McKim, Ernan McMullin, and Martha Nussbaum. Mike Loux was my first teacher of philosophy, and no one could ask for a better start in the subject—though given Mike's ability to galvanize students, "jump start" would be a more apt description. David Solomon has an incomparably encompassing view of big pictures in the history of philosophy, and he has tried—with less success than he might hope—to get me to see and communicate them. Alvin Plantinga and Tim Scanlon taught me that depth of understanding is best won through uncompromising rigor.

Some of my former teachers are now my colleagues; most are, or were, my friends. It is a privilege to dedicate this book to my teachers with the deepest gratitude for all that they have done for me.

List of Tables

This page intentionally left blank

Why Political Liberalism?

This page intentionally left blank

Introduction

In the 1980s, John Rawls—author of the magisterial work *A Theory of Justice*[1]—took what is sometimes described as a "political turn." Justice as fairness, the conception of justice presented in *TJ*, was re-presented as what Rawls called a "political liberalism." This re-presentation drew on a family of ideas and arguments that were new to justice as fairness, and reached its fullest expression in Rawls's second major work, *Political Liberalism*.[2] In this book, I take up the important but underexplored question of why Rawls made the turn to political liberalism. Answering this question has a number of textual and philosophical payoffs. One is that it leads us to a fuller appreciation of the deep problems that Rawls tried to address by developing a theory of justice.

An explanation of Rawls's turn to political liberalism should account for the differences between *TJ* and *PL*. Those differences are numerous and striking. I cannot discuss them all, and so it may help if I begin by listing those that I think stand in greatest need of explanation.

- In *PL*, the stability of a well-ordered society (WOS) is secured by an overlapping consensus of reasonable comprehensive doctrines.

1. John Rawls, *A Theory of Justice* (Harvard University Press, 1971 and 1999). I shall hereafter refer to this work as '*TJ*' and cite it parenthetically in the body of the text. The first page references are to the 1971 edition, and the second are to the revised edition of 1999.

2. John Rawls, *Political Liberalism* (Columbia University Press, 1996). I shall hereafter refer to this work as '*PL*' and cite it parenthetically in the body of the text.

- Justice as fairness is presented in that book as a political conception of justice, founded on basic ideas drawn from democratic political culture.
- In *PL*, the conception of the person represented by the original position—hereafter "the OP"—is said to be a political conception.
- The idea of public reasoning, which was hardly mentioned in *TJ*, is prominent in *PL*.
- The notion of political legitimacy, which received no explicit treatment in *TJ*, assumes a very prominent role in *PL*.
- In *PL*, Rawls admits that consensus in a WOS would probably focus on a family of liberal political conceptions of justice rather than on justice as fairness alone.
- *TJ*'s attempt to show that justice as fairness would be inherently stable is replaced in *PL* by an attempt to show that it would be stable "for the right reasons."

Three other changes are less obvious but very important: Rawls's description of the sense of justice and his argument that political society is a good undergo subtle but revealing changes, and the notion of congruence—so central to Rawls's treatment of stability in *TJ*—does very little work in *PL*.

These are the changes in Rawls's presentation of justice as fairness that I shall try to explain. Rawls made the changes to address shortcomings in the original presentation of his work. I take the position that Rawls thought the shortcomings he found were not merely shortcomings of interpretation, on his readers' part, but were shortcomings in justice as fairness itself and—in particular—in its treatment of the stability of a WOS.

I have tried to offer periodic summaries throughout the book, and have provided numerous of cross-references. I therefore hope that the book will prove easy enough to navigate that I need not supply a detailed map or summary at the outset. Instead, I shall confine myself to a few remarks that will, I hope, provide a useful overview of the journey to come. The best way to furnish that overview may be to communicate the surprise that readers of this book, or parts of it, have expressed about the picture of Rawls's work that emerges from it.

§1: Overview

A number of readers have said that the book introduces them to a very different Rawls than the one they thought they knew. Some of these readers still think of Rawls as a social choice theorist or a decision theorist. This book, they think, is not about the contractualist who once wrote that "the theory of justice is a part, perhaps the most significant part, of the theory of rational choice" (*TJ*, p. 16/15). Others have found this book surprising because they started with a quite different picture of Rawls. They think my claim that Rawls devoted considerable attention to avoiding collective action problems implies that he

was not the Kantian with whom they have become familiar. Collective action problems are indebted to one view of human rationality, they think, while Kantianism is animated by quite another. Still others have thought my argument that the justice of a well-ordered society depends upon large-scale changes in citizens' rational preferences shows that Rawls must be committed to a very non-Kantian account of moral motivation. Some readers have been surprised to meet a Rawls who has a persistent interest in the self and its unity. More have been surprised to meet a Rawls moved by deep questions about the goodness of humanity and the world.

These readers all started with something of the truth about justice as fairness. But as one reviewer of *TJ* said, "Rawls's theory has both the simplicity and the complexity of a Gothic cathedral."[3] These readers' surprise shows that they missed a great deal by adopting just "one view of [that] cathedral"[4] and by seeing Rawls's work from just one point of view. In this book, I try to develop and defend an interpretation that unifies their various perspectives and shows what truth there is in the various partial readings interpreters have extracted from Rawls's texts. I hope that the interpretation I defend is not only compelling, but also elegant and powerful in roughly the way that physical theories, economic theories, and mathematical results can be. Theories and results are elegant and powerful if they unify a lot on the basis of a little. I hope to do just that, showing how much of Rawls's work—including the most notable changes between *TJ* and *PL*—can be explained by supposing that he maintained a disciplined focus on a few intellectual concerns, and by seeing where those concerns led him.

One of Rawls's most pressing concerns was with the stability of a just society. He took up problems of stability in the third part of *Theory of Justice* and later in *Political Liberalism*. Seeing how Rawls initially thought he had shown that justice as fairness would be stable, and why he came to think that his original arguments for stability failed, shows why Rawls recast his view as a "political liberalism". By asking what Rawls means by 'stability' and what threats to stability he wanted to avert, we can unify the various perspectives on Rawls's work that I referred to a moment ago.

On my reading, Rawls wanted to identify basic terms of social cooperation that would be fair and collectively rational. Having identified those terms, he wanted to show that an arrangement which satisfied them would not be destabilized by a generalized prisoner's dilemma. At the same time, he wanted to show that they could be stabilized without reliance on a Hobbesian sovereign or a dominant ideology. Rather, he wanted the terms of cooperation to be

3. John Chapman, "Rawls's Theory of Justice," *American Political Science Review* 69, 2 (1975): 588-93, p. 588.

4. The phrase alludes to the title of Guido Calebresi and A. Douglas Melamed, "Property Rules, Liability Rules and Inalienability: One View of the Cathedral," *Harvard Law Review* 85, 6 (1972): pp. 1089–1128.

stabilized over time by the free activity of those who lived under them, in some robust sense of 'free'

Rawls argued in *TJ*, and continued to believe in *PL*, that justice as fairness would be stable only if citizens in a WOS developed a sense of justice. He argued that they would. He also thought that justice as fairness would remain stable only if citizens of a WOS maintained their sense of justice. Maintaining a sense of justice requires a commitment to leading a certain kind of life. *TJ*'s treatment of what Rawls called "congruence" was supposed to show that members of a WOS would affirm and maintain their commitment to living justly, so that their sense of justice would be a standing element of their character.

In his *Lectures on the History of Moral Philosophy*, Rawls says Kant believed that an enduring good will may require "a kind of conversion" that is "strengthened by the cultivation of the virtues and of the ways of thought and feeling that support them."[5] The religious overtones of the word 'conversion' open the possibility that Kant thought the maintenance of a good will is a response to supernatural intervention in one's life, a response that may need to be sustained by divine aid.[6] Despite his affinities with Kant, Rawls clearly wanted to furnish a naturalistic account of how members of a WOS sustain their good will, or that ingredient of a good will that stability requires: their sense of justice. His argument that members of a well-ordered society would maintain their sense of justice therefore relies on a naturalistic psychology and, in particular, on a tendency to reciprocity that was, he conjectures, naturally selected for.

Because of this important feature of human psychology, Rawls argued that the "ways of thought and feeling" that support a sense of justice can be fostered by just institutions. Such institutions would shape the characters of those who live under them, so that they would respond in kind to benefits received, and would attach little value to what they could gain from free-riding and other forms of injustice. Caring little about these gains, they would not be drawn to plans of life that would leave them free to decide case-by-case whether to honor the principles of justice. Instead, they would adopt plans that would give their desire to honor the principles a central place. Because each member of the WOS would adopt such a plan, and would know that everyone else would do so as well, justice as fairness would be stable. Because the character formation necessary for stability would be effected by institutions that satisfy the principles of justice, and because those principles are the centerpiece of justice as fairness, Rawls concluded that justice as fairness—when institutionalized and publicized—would stabilize itself.

5. John Rawls, *Lectures on the History of Moral Philosophy* (Harvard University Press, 2000), ed. Barbara Herman, p. 155.

6. Patrick Freierson, *Freedom and Anthropology in Kant's Moral Philosophy* (Cambridge University Press, 2003), p. 191, notes 31, 32, and 35.

Thus the Rawls of *TJ* recognized that an agreement reached in the original position could be undermined by a generalized prisoner's dilemma. Thinking he had shown that citizens of a just society would become the kind of persons who discount the pay-offs of injustice, he believed he had found a way to avert that threat without relying on a Hobbesian sovereign to alter citizens' pay-off tables. Furthermore, Rawls argued, because of the conditions of the original position, the principles that would be chosen there are principles members of the WOS would give themselves. And so when they regulated their lives by the principles, they would live lives that would be free in an important sense of 'free': they would live *autonomous* lives. Indeed, Rawls thought that one of the reasons they would endorse life-plans regulated by the demands of justice is that they would all want to live autonomously. Thus, *TJ*'s Kantianism was an essential part of Rawls' solution to the generalized prisoner's dilemma and his treatment of stability.

The possibility that members of the WOS would defect from fair terms of cooperation manifests a deep and familiar fact about human beings: we are creatures of divided hearts and wills. We can know what we should do and we can want to do it, but we can also be powerfully drawn to do something else—to advance our own interests, or those of people and causes we care about, in ways that are contrary to justice. This divide is a divide within our practical reason, a divide between what Rawls would come to call the Reasonable and the Rational. The stability of justice as fairness requires that our practical reason be unified and that our commitment to justice be—as Rawls would put it in *PL*—"wholehearted" (*PL*, p. xl). Because we are essentially reasoning beings, it requires that our selves be unified.

Few readers have recognized that *TJ*'s arguments for stability were intended to address the threat of a generalized prisoner's dilemma and to do so by showing how treating the principles of justice as regulative unifies human practical reason. If those arguments had succeeded, their success would have constituted a stunning philosophical achievement. Unfortunately, they did not. In the years following the publication of *TJ*, Rawls continued to accept his own earlier arguments that members of a WOS would develop a sense of justice, though in *PL* he made some important changes that he failed fully to acknowledge. But he came to realize that his argument that members of the WOS would maintain their sense of justice failed, and with it, his argument that a WOS would not be destabilized by a generalized prisoner's dilemma. And so he came to realize that he needed to offer a different set of arguments for those conclusions. Offering those new arguments required Rawls to recast justice as fairness as a political liberalism. The changes between *TJ* and *PL* that I listed above can be explained by seeing how they facilitate those new arguments.

Rawls's arguments for stability, both early and late, depend upon our natural amenability to developing a sense of justice and our natural amenability to the other developments of our character that just institutions are supposed to bring about. We can be naturally amenable to these developments

only if we have what the Rawls of *PL* called a "moral nature." By that he meant "not . . . a perfect such nature, yet one that can understand, act on and be sufficiently moved by a reasonable political conception of right and justice[.]" (*PL*, lxii) And so I believe Rawls thought that we can be amenable to the requisite moral development only if we are, or under the right circumstances can become, good. The arguments for stability in *PL*, if sound, vindicate the claim that we can be. If we are at least capable of being good, then—however we may actually behave—our presence in the world need not mar creation. The upshot, as I shall argue in the Conclusion, is that Rawls's theory of justice can be read as a brilliant and subtle exercise in naturalistic theodicy. Rawls offers arguments one consequence of which is that, despite the evil for which human beings are responsible, a good Creator could still have seen fit to fashion a world with us in it.

§2: The Road to Come

I have sketched my interpretation in broad strokes to provide readers some orientation, but the journey that follows goes by way of considerable textual and philosophical detail. According to the reading put forth here, Rawls took his political turn because there were clearly identifiable arguments in the original presentation of justice as fairness with which he later became dissatisfied. We can explain the changes between *TJ* and *PL* only by locating those arguments, laying them out with care, supplying missing premises when necessary, and asking where Rawls might have thought those arguments went wrong. We can then pinpoint key premises he came to reject as implausible, and others that he modified to facilitate his political turn.

I am not, of course, the only reader of Rawls who thinks we need to look at shortcomings of argument to find reasons for his political turn, but my reading of Rawls's reasons for the turn to political liberalism stands in sharp contrast to the interpretation that I think is most popular. That interpretation, which I call the *Public Basis View*, locates the shortcomings in an argument for the principles of justice that is said to be implicit in part I of *TJ*. That argument for the principles, which I call "the Pivotal Argument," is itself of considerable interest and serves as a useful analytic device to which I shall return periodically throughout the book. I therefore take some pains to lay it out precisely in Chapter I. Once the argument is laid out, the *Public Basis View* can be seen to have considerable appeal. I shall argue, however, that it founders on textual and philosophical shortcomings that prove insuperable.

I have said that the arguments with which Rawls became dissatisfied are to be found in the part of *TJ* devoted to the stability of justice as fairness and, in particular, in *TJ*'s treatment of congruence. In Chapter II, I distinguish various kinds of stability and identify the kind in which the Rawls of *TJ* was most interested—what he referred to as "inherent stability." Chapter II also

identifies, more clearly than is often done, the threat to stability with which Rawls was concerned. As I have already indicated, showing that justice as fairness would be inherently stable required showing that it could, when institutionalized, survive the threat of the generalized prisoner's dilemma without relying on a Hobbesian sovereign.

Chapters II and III show, in general terms, that *TJ*'s argument for the congruence of justice and goodness is a crucial part of Rawls's larger argument that justice as fairness would survive that threat, and so would be inherently stable. The problem with *TJ*'s treatment of stability, Rawls came to think, was that it relied on the improbable assumption that members of the WOS share what he called a "comprehensive doctrine." In Chapter III, I spell out what Rawls means by "a comprehensive doctrine," what he means by "congruence," where he thought his treatment of congruence relied on the assumption about a shared comprehensive doctrine that he later found implausible, and why reliance on that assumption in *TJ* led to an inconsistency in justice as fairness.

Some of the best published literature that treats of Rawls's congruence arguments mistake the structure of the congruence of arguments, the sequence of arguments that are offered, and the ways in which the various congruence arguments hang together. I give a good deal of attention to reconstructing those arguments, since I think we will see where Rawls thought the arguments went wrong only if we first see how he originally intended them to go. Chapter IV lays the groundwork for those arguments by attending to the acquisition of the desires they presuppose. Chapters V through VII lay out the arguments. In Chapter VIII, I go through the steps by which Rawls's treatment of congruence—so carefully knitted together in *TJ* and, as we shall see, in the original *Dewey Lectures*—came unraveled.

In Chapters IX and X, I show how the changes introduced between *TJ* and *PL* respond to the difficulties Rawls found in *TJ*'s treatment of stability. In the conclusion, I answer the question that gives this book its title by defending political liberalism against a common but powerful objection, by contrasting justice as fairness with another version of political liberalism, and by showing how political liberalism helps to answer the questions about the goodness of humanity and the world that, I have said, concerned Rawls so deeply.

§3: A Deeper Understanding of Justice as Fairness?

The congruence arguments in *TJ* are laid out in a single section late in the book. The claim that Rawls took his political turn because of problems in his original treatment of congruence might be thought to suggest the implausible thesis that Rawls made very far-reaching changes in his view because of shortcomings in a couple of pages of argument. In fact, as we shall see, the problems that Rawls identified in his treatment of congruence go to the heart of his

constructivism. That is one of the reasons Rawls came to think that the repairs needed by justice as fairness had to be so extensive. Moreover, the congruence arguments, when properly reconstructed, are seen to draw on material and concerns from throughout *TJ*. Making explicit how they did so brings some of the concerns and structure of *TJ* to light. One thing that is apparent from the recovery of the congruence arguments, for example, is that Rawls's concern with intuitionism—which he seemed to dispatch by the end of *TJ*, §7—was much more profound and pervasive than it is usually thought to be. Another is that the ambitious but puzzling discussion of the unity of the self in *TJ*, §85 responds to Rawls's deep and abiding concerns about how practical reason is to be unified. Appreciating that section, I believe, deepens our appreciation of the Kantian Interpretation of justice as fairness laid out in *TJ*, §40. It especially heightens our appreciation for the crucial role Rawls assigned a Kantian conception of the person in *TJ*.

As these remarks suggest, one striking feature of the treatment of congruence is the extent to which it draws on other sections of *TJ*, and on other sections of part III in particular. One of the reasons we learn so much about *TJ*, and about justice as fairness, by asking why Rawls turned to political liberalism is that we come to see how parts of *TJ* fit together, in unanticipated ways, by making the congruence arguments explicit. Part III of *TJ* is sometimes read as if it were an undisciplined attempt to cover some of Rawls's favorite topics in ethics. The material on the moral and natural sentiments, for example, can appear to be set of tangential arguments directed against crude forms of emotivism and prescriptivism. In fact, I believe part III is exemplary for the way it painstakingly establishes conclusions with an eye toward their later use in Rawls's arguments for stability. We shall see that the continuity of the sentiments is crucial for the second congruence argument Rawls offers in *TJ*, §86. While this book is not a commentary on part III of *TJ*, I hope it will go some way toward rekindling interest in that neglected part of the book.

Pursuing the reasons for Rawls's political turn also puts us in a position to see how much of the treatment of moral development in *TJ*, chapter 8 survives the transition to *PL*. This is a natural question to raise about justice as fairness, since Rawls rarely spoke of a sense of justice after *TJ* and did not return to the process of moral development in any systematic way. But I do not think that that is because other matters eclipsed his concern with the development of a sense of justice or because he thought his discussion of moral development needed to be abandoned. Rather, as I hinted earlier, Rawls continued to think the question of whether a WOS would be stable had a two-part answer. The first part was provided by showing that members of the WOS would develop a sense of justice. The second was provided by showing that they would judge that preserving their sense of justice belongs to their good. Rawls did not revisit *TJ*'s treatment of the first part in subsequent work because, he says, he continued to think it was adequate, and could survive the changes in his view. Rawls made the changes between *TJ* and *PL* because he thought they were necessary to support the second part of the answer; I shall suggest that, his

claims to the contrary notwithstanding, Rawls himself thought changes in the first part—at least changes of emphasis—were called for as well.

An especially important question about justice as fairness concerns the dispensability of the original position. That question has hung over Rawls's work for almost four decades. Rawls's insistence in his later work that the original position is a device of representation seems to invite the question in urgent form, but that question was pressed in some quarters well before the political turn. I argue that the original position is a theoretical device that "bridges" the right and the good in Rawls's early work, for it functions in the argument by which Rawls identifies principles of right and in an argument by which Rawls argues that acting from those principles belongs to the good of members of the WOS. The original position may not be necessary for the first argument but, I shall argue, it is necessary for the second. The second argument was, in turn, necessary to solve the question of congruence in *TJ* and the *Dewey Lectures.* The original position is not, therefore, dispensable from the arguments Rawls offered for justice as fairness before his political turn.

Perhaps the most notable feature of Rawls's re-presentation of justice as fairness is its starting point. Rawls insists that as a political liberalism, justice as fairness begins with ideas and convictions latent in the public political culture of liberal democracy. Most readers have considered this to be a marked—if not a revolutionary—change from the philosophical method of *TJ*. Some, as we shall see, have accused Rawls of moral retrenchment. I shall argue, against the conventional wisdom, that even in *TJ*, Rawls took for granted a view that members of liberal democratic societies can normally be expected to have of themselves, and that in the course of developing justice as fairness he refined that view of the person and gave it a central role.

Thus even before his political turn, Rawls started from within—and addressed his work to—the liberal democratic world. The difference between his earlier and later presentations of justice as fairness is not, therefore, that the latter starts within that world while the former does not. The difference lies in what he drew from liberal democratic culture. In his early work, it was an ethical—not a metaphysical—conception of the person, a conception that he further specified in ways that he came to think could be an object of controversy among reasonable citizens. In his later work, he was made clear that the conception of the person he drew from political culture was a specifically political conception.

§4: Unity, Theodicy, and the Attractions of Liberalism

By looking closely into why Rawls made the changes between *TJ* and *PL*, we also learn a great deal about liberalism, its attractions, and its ambitions.

The theoretical foundation of liberalism is sometimes said to be a set of rights or a basic right, such as the right to equal concern and respect. That is

why some readers, most famously Ronald Dworkin, interpret Rawls's liberalism as rights-based. Though Charles Larmore has argued that a principle of legitimacy lies at the core of political liberalism, he thinks that what the principle of legitimacy really expresses is an imperative of respect for persons, and so his reading has strong affinities with Dworkin's.[7]

The role of reflective equilibrium in justifying justice as fairness implies that there is some artificiality to speaking of a "foundation" for Rawls's liberalism. Those qualifications notwithstanding, the reading of Rawls that I defend here shows that justice as fairness is an alternative to rights-based—and hence to legitimacy-based—theories of justice. On my reading, Rawls supposes from the outset that under the impact of liberal democratic thought and practice, we, his readers, think of ourselves as free and equal persons embedded in a society that ought to be a fair scheme of social cooperation. We have, he thinks, a democratic conception of our society and a conception of ourselves that I call a *free-and-equal self-conception*.

Crudely put, Rawls refines and specifies these conceptions so that they yield an answer to the question he poses in the *Dewey Lectures*: what conception of justice is best suited to regulate the collective political life of persons who think of themselves as free and equal members of a fair cooperative scheme? Liberal rights, and a liberal conception of legitimacy, are not the foundations of his liberalism, though they are part of Rawls's answer to that question. As we shall see, his principle of legitimacy, as stated in *PL*, is justified by showing that our exercises of political power must conform to that principle if we are to live as free and equal persons, properly conceived, and to enjoy what I shall call the *Ideal of Democratic Governance*. Thus, if we can speak of the "foundation" or "foundations" of justice as fairness at all, what is foundational to it are conceptions of the person and of society that are found in democratic culture and that are made specific enough to generate political principles. Justice as fairness therefore illustrates—as Rawls himself says—the possibility of a liberalism that is "conception-based" or "ideal-based," rather than "rights-based."[8]

The attraction of Rawls's principles of justice depends in part upon their distributive implications. But it also depends on the attractiveness of the political conception or ideal of the person on which they are based, for among the reasons we have for acting from the principles is that by doing so, we will realize that ideal. That ideal is, I believe, very attractive. Its attractiveness is important. Some critics, put off by what they see as the individualism, selfishness, and materialism of modern life, claim that liberalism invariably

7. See Charles Larmore, "The Moral Basis of Political Liberalism," in Larmore, *The Autonomy of Morality* (Cambridge University Press, 2008), pp. 139–67, especially pp. 146ff.

8. John Rawls, "Justice as Fairness: Political not Metaphysical," in John Rawls, *Collected Papers*, (Harvard University Press, 1999), ed. Samuel Freeman, pp. 388–414, pp. 400–401, note 19; Rawls credits Elizabeth Anderson with describing his view as "ideal-based."

produces the kind of person they deplore. They defend other forms of political life as better suited to our social nature. One way to answer these critics is to show that liberalism does take due account of our social nature, and encourages us to live up to conceptions of ourselves that lack the features on which critics seize.

Rawls's liberalism suggests how this might be done. Rawls is often read as propounding an individualistic theory. The argument for the principles, which relies on the device of a social contract, can be described that way. But according to *TJ*'s arguments for congruence, members of the WOS would judge that upholding the principles is part of their good because it is only by upholding the principles that they can satisfy natural desires for friendship, association, and sincere and open dealings with others. Though Rawls modified those arguments considerably in his later work, he continued to think that part of what makes his principles attractive is that acting from them enables us to live among others in ways that should appeal and inspire.

Some readers have said that on reading *TJ*, they thought that their own deepest moral convictions had received their best expression and their most powerful defense. Others of us had a somewhat different reaction. Justice as fairness expressed our deepest political convictions. But we came to political philosophy with deeply held views about what is good in life and why, and those conceptions of the good had implications for the right that were not obviously compatible with justice as fairness. The result was a tension between potentially conflicting identities.

In Rawlsian terms this tension reflects a conflict between the demands of conceptions of justice associated with our views of the good, on the one hand, and the demands of the Reasonable on the other. The attraction of justice as fairness is not, therefore, the attraction of something that is alien to those who have traditional views of the good. It is the attraction we feel for the reasonable part of ourselves. Rawls's concern with the unity of the self showed the tremendous ambition of *TJ* and promised to show how the the tension should be resolved. For Rawls argued that the only way creatures like us can live as unified selves, at least under modern conditions, is to regulate our pursuit of the good by principles of liberal democratic justice. The alternative to being regulated by the reasonable part of ourselves was, Rawls seemed to suggest, to live lives that lacked rational unity. That is why—though Rawls had said of the parties in the OP that their aim "is to establish just and favorable conditions for each to fashion his own unity" (*TJ*, p. 563/493)—he also said that what he called the "essential unity" of the self is established by taking the principles of justice as supremely regulative (*TJ*, p. 563/493).

An important part of the congruence argument, I will suggest, is devoted to establishing this last claim. We shall see that one of the reasons Rawls became dissatisfied with his treatment of congruence was that he realized a truly liberal view cannot take a stand on how the "essential unity" of selves is to be attained. And so while he continued to think that each citizen in the WOS would treat the principles of justice as in some sense regulative, he also

came to recognize that how the principles of justice are to be connected with or founded on various conceptions of the good must be left to each person to work out. I believe that one reason for taking Rawls's principles as regulative of our political lives is the great attraction of being the kind of citizens justice as fairness calls us to be. Seeing that we can be that kind of citizen, in turn, completes what I referred to earlier as Rawls's "naturalistic theodicy," for it vindicates our hope in the possibility of a world that is more just and that can aptly be called "good."

§5: A Final Word to the Reader

I have given some indication of what I think can be learned by pursuing questions about why Rawls made the changes he did between *TJ* and *PL*. I conclude this introduction by saying a few words about what I shall ask of readers and about the limitations of the book.

As my remarks so far have suggested, this book is not intended as a primer in the main lines of Rawls's thought. Moreover, at this point, the literature on Rawls is so well developed, and the study of his work so widespread and thorough, that I feel justified in presupposing an acquaintance with the major ideas and texts that is fairly sophisticated. A sign of the familiarity that I presuppose is that I use abbreviations like WOS for "well-ordered society" and OP for "original position." Because Rawls's texts and ideas have attracted so much critical attention, I also assume that any reading that hopes to offer something new must be very carefully defended and very firmly anchored in the text. I have therefore hewn closely to the written word and used an expository style that is more commonly found in other areas of philosophy, spelling out some of Rawls's reasoning in premise-and-conclusion form. Some of Rawls's arguments compress very complicated lines of thought and, as I have already implied, the compression in *TJ* is facilitated by Rawls's frequent reliance in one argument on conclusions that have been established by other arguments elsewhere in the book. The reconstructions that I provide can therefore be demanding. I have made demands of readers because I believe the reconstructions heighten appreciation for the rigor of Rawls's own arguments, and that the method of exposition I have chosen makes analysis of those arguments more economical and perspicuous.

Some of the most demanding reconstructions are in Chapters IV through VII, where *TJ*'s congruence arguments are laid out and analyzed. Chapter VIII, which tells why Rawls became dissatisfied with those arguments, depends upon the chapters that immediately precede it. These four chapters together supply the interpretation offered here with some of its most detailed textual and philosophical support. As I have already indicated, Chapters II and III provide an overview of *TJ*'s treatment of stability and of the reasons Rawls became dissatisfied with it. Readers who are less interested in the details of the

congruence arguments, who are uninterested in textual exegesis, or who are content with a general understanding of why Rawls made the turn to political liberalism, are invited to read selectively between Chapter III and Chapters IX and X. There, I show how the changes Rawls made after *TJ* respond to the sources of his dissatisfaction with his earlier arguments.

This book is intended to be a defense of political liberalism, but it is a defense of an unusual kind. Though I do reply to some standard objections to political liberalism in the Conclusion, the book is not an attempt to defend Rawls's later views against all comers. Rather, the defense provided here is the kind of defense Gerald Cohen hoped to provide of Karl Marx's theory of history—a defense that proceeds "by offering argument in its favor, but more by presenting the theory in what I hope is an attractive form."[9] While I did not face the challenge that Cohen did, I thought that one attractive form in which political liberalism still needed to be presented is as a rigorous and systematic response to a specific set of problems which Rawls correctly came to see in premises and arguments on which he had previously relied. I hope that my end is served by the care with which I have tried to lay out Rawls's lines of thought, both early and late, and by my attempt to display the underlying unity of his views.

I am strongly inclined to think that Rawls succeeded at what he set out to do: identify fair and collectively rational principles of justice that, when institutionalized and publicized, avert the threats to stability with which I have claimed he was concerned. Unfortunately, laying out and unifying Rawls's treatment of stability within tolerable bounds of length meant giving less critical scrutiny to certain crucial claims than I would have liked. There are many places at which what Rawls says admits of more than one interpretation, at least when what he says is taken in isolation. Quite often, I have assumed that readers of Rawls will already have noticed the ambiguity and that my job is to stake out a position on an interpretive question rather than to belabor the way the question arises. In these cases, I have opted for what I take to be the best reading and shown that it makes sense of the larger argument, without explicitly distinguishing and puzzling through the various interpretations the text will bear.

As I have indicated, Rawls's arguments for stability depend upon psychological assumptions. Those assumptions need probing. One assumption, or set of assumptions, is especially in need of attention: Rawls's assumption that acquisitiveness has its origins in the desire for status. This assumption does considerable philosophical and political work in justice as fairness. It is an assumption Rawls held throughout his working life.[10] In §V.4, I have tried

9. Gerald Cohen, *Karl Marx's Theory of History: A Defence* (Princeton University Press, 1978), p. ix.

10. See my review of John Rawls, *A Brief Inquiry into the Meaning of Sin and Faith* (Harvard University Press, 2009), ed. Nagel, *Notre Dame Philosophical Reviews*, http://ndpr.nd.edu/review.cfm?id = 17045

to understand why the assumption might hold in a special case and I have expressed some skepticism about it elsewhere,[11] but I have not subjected it to anything like the attention it deserves. Unfortunately, that will have to await another occasion. My aim has been to convey a synoptic view of how and why Rawls rebuilt his cathedral; doing so left me less scope than I would have liked to test this particular buttress.

Academic work is a way of serving others. I recognize that this book may be of greatest service to those who have wrestled with Rawls's texts for a long time, who remain puzzled about how certain of his arguments go and who wonder what he could have meant by certain obviously crucial but vexing assertions and turns of phrase. Even after some decades of scholarly attention to Rawls's work, I believe there is still a need for a book that pays attends so closely to texts that bear on his political turn and that tries to figure out exactly how his arguments go. That is the need I have tried to fill here. But I hope that this book will also be of service to all those who wonder whether a just world is possible, whether we human beings are capable of sustaining such a world, and whether those of us with traditional conceptions of the good can achieve some unity of self while living with others as free equals under modern conditions. These questions were, I believe, of the deepest concern to the greatest political philosopher of our time. In writing this book, I have tried to understand how he posed and answered them.

11. In my review of Rawls, *Brief Inquiry*; also in my "John Rawls and the Task of Political Philosophy," *The Review of Politics* 71 (2009): pp. 113–25.

I

The *Public Basis View*

Rawls made the changes between *TJ* and *PL* because he became dissatisfied with arguments that were critical to the presentation of justice as fairness in his first book. Any serious attempt to explain those changes must therefore identify the arguments with which Rawls became dissatisfied and say why he came to think they were unsatisfactory. In Chapter II, I shall say what I think those arguments were and what problems Rawls found with them. My account of the changes between *TJ* and *PL* challenges what I take to be the standard explanation of those changes. I shall refer to that explanation as the *Public Basis View* of the changes, and I shall devote this chapter to laying it out and evaluating it.

The label I have attached to the *Public Basis View* is new, but I think the *View* itself is widely accepted. Indeed, I believe that most readers who have an opinion about why Rawls introduced the changes between *TJ* and *PL* accept the *Public Basis View* in some form. I shall begin by developing the *Public Basis View* of the changes as an ideal type. I believe that the essentials of the *View* will be recognizable to those familiar with literature about, and discussion of, Rawls's turn to political liberalism. Later, I shall suggest that some philosophers who have developed prominent political liberalisms of their own endorse the *Public Basis View* of Rawls's political turn.

§I.1: Initial Statement of the *Public Basis View*

The *Public Basis View* of Rawls's transition is most easily explained and made vivid by relying on a certain picture of Rawls's WOS—a picture according to

which the WOS has a public charter that is expressed in fundamental political documents which play roughly the role in that society that the Declaration of Independence and the Constitution play in American political culture. While Rawls himself may not have had that picture in mind, it is not out of the question that he did and, as I hope will be evident, the picture has some heuristic value.

Rawls says that in the WOS of *TJ*, everyone would accept and would know that everyone else accepts the same conception of justice—just as, in the United States, citizens recognize and know that others recognize the rights and liberties accorded everyone by the Constitution. It is that conception, Rawls says, that serves as the WOS's "foundation charter" (*TJ*, p. 11/10). By that Rawls meant that it was to serve as the shared, public basis for distributing benefits and burdens of social cooperation. If justice as fairness were to serve as a *shared* basis of justification, then it would have to be defended with an argument or a set of arguments that could be affirmed by all members of the WOS, so that everyone would accept the same principles of justice and accept them on the same grounds. This is the sort of defense Rawls hoped to provide in part I of *TJ*.

In the WOS of justice as fairness, the defense of the principles would be publicly available in important documents, just as the philosophical justification of American government is alluded to in the Declaration of Independence. That justification is alluded to in the second paragraph of the Declaration, which famously begins:

> We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.—That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed[.]

Thus, according to the publicly available foundation of the American government, the ends, limits, and powers of government are justified—via consent—by a conception of the person. Similarly, according to the *Public Basis View*, the publicly available justification of justice as fairness would justify *it*—via consent—by a metaphysical conception of the person. In the Founders' United States, the publicly articulated, metaphysical conception of human beings asserts that we are created free and equal by God. So in the WOS of *TJ*, it might be thought, the publicly articulated metaphysical view of human beings would be or would seem to be the Kantian view of human autonomy and equality expressed in the original position.[1]

According to the *Public Basis View*, the problem Rawls came to see grows out of the fact that the WOS of justice as fairness would be a liberal society. Its

1. For Rawls's exposition of the Kantian interpretation of justice as fairness, see Rawls, *TJ*, §40.

members would be free to explore and adopt a variety of religious and philosophical views about the good—what Rawls later labeled "comprehensive views" or "comprehensive philosophical doctrines" (*PL*, p. xviii). As Rawls came more deeply to appreciate this "fact of reasonable pluralism" (*PL*, p. 36), he came to see that the Kantian conception of the person is not a neutral starting point for political theorizing, but is a conception with which many reasonable people in a pluralistic society would disagree. At the same time, it is said, critics like Michael Sandel showed just how heavily the original position argument for the two principles depended upon the contentious metaphysical conception of the person from which Rawls began. Rawls therefore realized that reasonable people in a pluralistic society might reject the metaphysical argument he provided—or could be read as providing—for his principles. Since the WOS of justice as fairness would be a pluralistic society of reasonable people, he came to realize that the WOS might not be one in which everyone accepted the same conception of justice and its public defense after all.

To remedy this tension or seeming tension in his view, proponents of the *Public Basis View* claim, Rawls recast his defense of the principles so that it rested on premises that could be accepted by citizens who adhered to a wide variety of conceptions of the good and of the person—premises that were compatible with those conceptions because they were "political not metaphysical." The public defense of justice as fairness was then explicitly said to begin, not from a metaphysical conception of the person, but from the conception of the citizen found in the public political culture of a democratic society. The principles of justice were then said to be justified—via consent—by this political conception of the person. The political premises of the new defense could then serve as the shared, public basis of the principles that Rawls had hoped to provide in *TJ*. Because members of the WOS endorse those premises from within their own comprehensive doctrines, the "foundation charter" of the WOS is, as it were, an area of "overlap" among otherwise divergent doctrines—hence the image of an "overlapping consensus."

This brief summary of the *Public Basis View* may exaggerate—or may draw out at greater length than any proponent of the *View* would—the parallels between the Declaration of Independence and the public defense of justice as fairness in the WOS. But by doing so, it makes vivid three of the central claims of the *Public Basis View*: (i) the claim that the argument with which Rawls became dissatisfied was the argument for the principles of justice provided in part I of *TJ*, (ii) the claim that Rawls became dissatisfied with it because he recognized that it would be too controversial to serve as the shared, public basis of the principles in a pluralistic society, and (iii) the claim that Rawls responded to this difficulty by recasting that defense so that it could be the object of an overlapping consensus.

In one respect, however, the summary is too simple, since it suggests that there is a single *Public Basis View*. But at a critical juncture in the summary, I said that according to the *Public Basis View*, Rawls came to realize that he "provided – *or could be read as providing*" a defense of his principles that relied

upon a metaphysical conception of the person. This disjunction suggests two different reasons for Rawls's dissatisfaction with *TJ*'s defense of the principles of justice. There are therefore two different versions of the *Public Basis View*, which I shall refer to as the "strong" and "weak" versions.

Proponents of the strong version claim that Rawls's defense of the principles of justice really did rely upon metaphysical claims about persons. In moving from *TJ* to *PL*, they say, he disavowed those claims in favor of other arguments for the principles, arguments the premises of which are "political not metaphysical." Thus, in its strongest form, the *Public Basis View* is a thesis about substantive changes in justice as fairness, which involve the rejection of some metaphysical claims that Rawls previously endorsed. It is now widely thought that the central contention of this version is mistaken, for *TJ*'s argument for the principles of justice is now thought not to depend upon metaphysical claims. Even if this is so, there are two reasons why the strong version of the *Public Basis View* remains worthy of attention. One is that it is instructive to see just what is meant by denying that Rawls relies on metaphysical claims, since—though this is not generally appreciated—I think Rawls himself had something fairly precise in mind in denying it. The other is that the failure of the strong version of the *Public Basis View* suggests the weaker—and hence the more broadly appealing—version of the *View*.

Proponents of the weak variant recognize that many readers of *TJ*—including proponents of the strong variant—took Rawls's defense of the principles to depend upon metaphysical assumptions. But they deny that Rawls ever meant the premises of his defense to be taken this way. They think Rawls took an explicitly political turn in order to make clear that this metaphysical reading of those premises was wrong. The new ideas introduced in *PL*—such as the ideas of an overlapping consensus, the political conception of the person, and political autonomy—are said to be ideas Rawls introduced to explain what he meant all along.

At the heart of the both versions of the *Public Basis View* is, of course, the argument for the principles with which Rawls is alleged to have become dissatisfied—because it either relied on metaphysical claims or seemed to rely on them. I shall offer a concise version of that argument in the next section. Since that argument is, as it were, the pivot around which he is said to have made his political turn, I shall refer to that argument as the "Pivotal Argument." In order to see the appeal—and what I shall argue are the fatal textual and philosophical shortcomings—of the *Public Basis View*, it is necessary to go beyond the rough statement of the *View* I have given in this section and to lay out that argument rigorously. Some of the steps are unfortunately rather cumbersome, but having the argument before us will make for economy and clarity later on, since I shall refer to some of the steps frequently in the chapters to come. I shall not contend that *Public Basis View* is mistaken in supposing that Rawls relied on the Pivotal Argument or on an argument very like it, nor shall I deny that Rawls modified certain key claims in the argument as part of his transition to political liberalism. About these things, the *Public Basis View* is importantly

right. What I do deny is that the Pivotal Argument is the argument with which Rawls primarily became dissatisfied after publishing *TJ*. The changes Rawls made in his defense of the principles were motivated by his dissatisfaction with—and his need fundamentally to rethink—a very different set of arguments, found in a different part of *TJ*.

§I.2: The Pivotal Argument

I said earlier that I am developing the *Public Basis View* as an idealized position with which to contrast my own explanation of Rawls's political turn. The Pivotal Argument is not, therefore, an argument that is explicitly attributed to Rawls in any one article of scholarly literature. Rather, it is an argument that has to be supplied as part of the rational reconstruction of a view about changes between *TJ* and *PL* that is widely, if implicitly, held. In this section, I attempt to supply it.

When I sketched the *Public Basis View* in the last section, I implied that the Pivotal Argument follows a sequence of thought that begins with an assertion about human nature and proceeds, via consent in the original position, to Rawls's two principles. What I have called "the Pivotal Argument" therefore begins with a claim about human nature:

(1.1) We are by nature free and equal rational agents who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

Rawls assumes that human beings need access to the primary goods regardless of what ends they adopt. Those goods are produced and distributed by the basic structure of society. Because access to these goods is necessary—and because these goods are distributed by institutions whose influence is pervasive—our life prospects, our aspirations, and our sense of what is just and unjust, all are deeply affected by the distribution of primary goods. This gives us a powerful interest in how primary goods are distributed. And so the second step in the Pivotal Argument is:

(1.2) We have a fundamental interest in the ways the basic structure of our society distributes the primary goods.

The fundamental interest we have in the production and distribution of primary goods makes their production and distribution a matter of justice. And so the principles in accord with which the basic structure produces and distributes primary goods must conform to what justice demands.

The task of determining what justice demands of the basic structure is, of course, the task Rawls sets himself in *TJ*. He locates his attempt to answer that question squarely in the contract tradition. Like others in the contract tradition, Rawls seems to make a crucial assumption about how basic social

arrangements are determined. Crudely put, forcing people to live under arrangements that are not acceptable to them is inconsistent with respecting them as the kinds of beings (1.1) says they are. More precisely:

(1.3) If we have a fundamental interest in basic social arrangements, and if we are capable of rationally assessing those arrangements in light of our interests, then respect for us as free and equal persons with that interest and capability requires that the principles governing those arrangements be acceptable to us as such persons.

(1.3) is a conditional. The consequent is conditional on the claim that persons have any fundamental interest in basic social arrangements at all. One such interest they would have is an interest asserted in (1.2), the interest in how the basic structure produces and distributes primary goods. So (1.3) seems to imply that:

(1.4) If (1.2) is true, and if we are capable of rationally assessing the ways the basic structure distributes primary goods in light of our interests, then respect for us as free and equal persons with that interest and capability requires that the principles governing the basic structure be acceptable to us as such persons.

I have already argued for (1.2). And (1.1) implies that we are capable of rationally assessing the way the basic structure produces and distributes primary goods. So (1.1), (1.2), and (1.4) imply:

(1.5) Our society respects us as the kind of persons (1.1) says we are only if the principles governing the ways the basic structure of our society distributes primary goods are acceptable to us as such persons.

If Rawls also assumes that persons must be respected by their society as the kind of being (1.1) says they are then, since (1.1) says we are free and equal persons, the assumption that we must be respected—together with (1.5)—implies that:

(1.6) The principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal persons.

What does it mean to say that principles are or are not *acceptable* to us? And what does it mean to say that they are or are not acceptable to us *as free and equal persons*?

To say that principles are acceptable to us is to say that, if given the choice, we would accept them. To say that principles are acceptable to us *as free and equal persons* qualifies or elucidates the conditions under which they must be accepted. A crucial move in the Pivotal Argument is the claim that if the principles that govern distribution among persons were determined by features of their situation that are irrelevant from a moral point of view, then those persons would not really be treated as equals, since equal treatment requires leaving such considerations aside. This assumption requires that those who choose

or accept the principles must determine the principles free of the influence of those contingencies. And so:

(1.7) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation that is uninfluenced by natural and social contingencies.

Once these contingencies are screened out, what is decisive in determining what principles we would accept is our nature as persons. There is nothing else left to determine the choice. So (1.7) implies:

(1.8) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal persons is the decisive determining element of the choice.

The first premise of the Pivotal Argument, (1.1), is a claim about what we are by nature: free and equal rational agents capable of reflecting on ends and assessing social arrangements in light of our interests. This is just the way that we are represented in Rawls's choice situation, the OP. Indeed, it seems, the OP is constructed precisely so that nothing other than our nature as described in (1.1) affects what principles are adopted there. So Rawls seems to think that:

(1.9) The OP is a choice situation in which our nature is the decisive determining element.

From (1.8) and (1.9), it follows that

(1.10) The principles governing the ways the basic structure distributes primary goods must be acceptable in the OP.

Acceptability in the OP is determined by a series of pair-wise comparisons. And since Rawls argues that his two principles would be chosen in preference to other principles in the OP, he concludes that:

C_1: The distribution of primary goods by the basic structure must be governed by the two principles.

This is the Pivotal Argument. It is the line of thought by which the *Public Basis View* alleges that Rawls's principles would be publicly justified in the WOS of *TJ*. It is also the line of thought with which readers sympathetic to the *View* allege that Rawls became dissatisfied.

§I.3: Imputing the Pivotal Argument?

The plausibility of the *Public Basis View* depends upon the plausibility of imputing the Pivotal Argument to Rawls. There are some textual bases for imputing it.

Some of those bases were canvassed by Michael Sandel. Sandel famously went to some lengths to argue that Rawls defended his principles of justice by

relying on claims about persons that Sandel interprets as metaphysical.[2] Sandel was undoubtedly right that there is a conception of the person at work in *TJ* according to which members of the WOS are as (1.1) describes them. Sandel was also right to claim that the work done by that conception includes shaping the OP. For in the original edition of *TJ*, Rawls says that "the desire for liberty is the chief regulative interest that the parties must suppose they will all have in common in due course" and that the veil of ignorance "lead[s] to this conclusion"[3] (*TJ*, p. 543). Since the principles of justice are defended by showing that they would be chosen in the OP, these remarks suggest that Rawls *did* rely on (1.1) or on some premise quite like it in *TJ*, and that he relied on it to defend the principles.

Furthermore, some of the crucial assumptions that underpin the Pivotal Argument—such as those made in the moves to steps (1.3), (1.6), (1.7), and (1.9)—seem to be assumptions on which Rawls relied. (1.3) expresses a quintessentially contractualist idea about what respect for persons requires. In moving from (1.5) to (1.6), the Argument assumes it is imperative to respect persons as the kind of being (1.1) says they are. This is an imperative Rawls is widely read as presupposing and, indeed, reliance on it may seem to be the source of much of his view's appeal. The step from (1.6) to (1.7) is taken on the basis of a claim Rawls seems to make explicitly, when he says that principles which are adopted without "exploitation of the contingencies of nature and social circumstance" express respect for those who live under them (*TJ*, p. 179/156). As we shall see later, (1.9) is necessary to sustain the Kantian Interpretation of justice as fairness (cf. *TJ*, p. 252/222).

But the Pivotal Argument is not one that Rawls ever lays out systematically nor can it be extracted from any one passage of *TJ*. This may engender some doubts about the claim that Rawls relies on it or any argument like it, and so may raise doubts about whether there is any plausible reading of Rawls that gives it a central place. These doubts may be heightened by two clearly identifiable ways in which the Pivotal Argument diverges from *TJ*'s defenses of the principles of justice, for if the Pivotal Argument omits considerations or arguments on which those defenses draw, then the *Public Basis View*'s claim to identify the sources of Rawls's dissatisfaction with those defenses would be undermined.

One especially notable and surprising departure from Rawls's texts seems to be that the Pivotal Argument accords the OP only derivative force in support of the principles of justice: the OP is not referred to explicitly until (1.9) and the argument does not go through the details of the parties' choice there. On the contrary, I think the secondary role given the OP tells in favor of imputing the

2. See Michael Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982), pp. 48ff., 133–34, and 175ff.

3. I take this passage to support the claim about freedom and equality, and not just freedom, because I take Rawls to mean that parties assume they all will have an equal interest in liberty in due course.

Pivotal Argument to Rawls rather than against it. For this reason, and because the objection raises deep issues that I shall take up later, I want to confront it.

In imputing the Pivotal Argument to Rawls, the *Public Basis View* builds on an insight that was first articulated by Ronald Dworkin. That insight is that in *TJ*, Rawls argues "through" the OP from more fundamental presuppositions.[4] Describing how he thinks Rawls argues through the OP, Dworkin writes:

> The original position is well designed to enforce the abstract right to equal concern and respect, which must be understood to be the fundamental concept of Rawls's deep theory.[5]

The Pivotal Argument seems to spell out Dworkin's insight by showing that Rawls argues through the OP in just this way. For at (1.10), the OP seems to do the enforcing to which Dworkin refers. The transition from (1.5) to (1.6) seems to depend on the right to respect that Dworkin says it enforces.

Dworkin's reading of Rawls is open to question. Moreover, there remains some controversy about just what Dworkin has shown even if his interpretation is right. That controversy bears on the plausibility of the *Public Basis View* and of other views, like my own, that attribute something like the Pivotal Argument to Rawls. Dworkin is sometimes thought to have shown, not just that Rawls argues through the OP, but that the OP is therefore dispensable. If this reading of Dworkin were correct, and if Dworkin's reading of Rawls is correct, then that would tell against imputing the Pivotal Argument to Rawls since the Pivotal Argument goes through the OP, but does not dispense with it. But this reading of Dworkin is a mistake. Dworkin argues that the OP does not have fundamental justificatory force. As I shall explain in §VII.9, nothing he says entails that it is dispensable. So Dworkin's reading does not imply that the Pivotal Argument should not be imputed to Rawls.

Someone working in the spirit of Dworkin *could* show that the OP is a dispensable part of the argument for C_1—the claim that primary goods must be distributed in accordance with the principles of justice—by producing an argument for C_1 that begins from the requirement of equal concern and respect but does not go by way of the OP. As we shall see in Chapter VII, such an argument in effect moves from (1.6) to C_1 differently than the Pivotal Argument does, by attaching a different interpretation to (1.6)'s requirement that principles be acceptable to us "as free and equal persons." Dworkin himself does not provide such an argument, but Joshua Cohen does.[6] In an important paper called "Democratic Equality," Cohen argues for Rawls's principles

4. Ronald Dworkin, "The Original Position," in *Reading Rawls* (Oxford: Basil Blackwell, 1975), ed. Norman Daniels.

5. Dworkin, "The Original Position," *Reading Rawls*, p. 181.

6. Joshua Cohen, "Democratic Equality," *Ethics* 99 (1989): pp. 727–51. Cohen does not cite Dworkin, and I do not mean to suggest that Cohen himself accepts Dworkin's interpretation of Rawls.

from the claim that principles must be acceptable to every social position. If we think that people have a right to equal concern and respect in design of institutions if and only if those institutions must be acceptable to each social position, then Cohen has outlined an argument for the principles that, in effect, begins with the requirement of equal concern and respect and justifies Rawls's two principles while bypassing the OP.

Of course, it does not follow from Cohen's argument—nor does Cohen say it follows—that the OP is dispensable altogether. Whether it is depends, as Cohen recognizes, upon whether the OP plays any essential role elsewhere in Rawls's theory of justice. I shall argue in §VII.9 that the OP is not dispensable from the theory as laid out in *TJ*, even if there are good arguments for C_1 that bypass it. What matters for present purposes is this. The fact that the Pivotal Argument gives the OP derivative force does not tell against imputing that argument to Rawls. Those who find Dworkin's reading of Rawls persuasive may think that the Pivotal Argument—or something like it—is needed to spell out the central reasoning of what Dworkin calls Rawls's "deep theory."

There may seem to be a second, more serious difficulty with imputing the Pivotal Argument to Rawls. According to Rawls, a WOS is to be, as he famously says, "a fair scheme of social cooperation." A scheme of *cooperation* is one conducted on terms that are mutually acknowledged. Rawls clearly thinks that the principles of justice are such terms. Yet the Pivotal Argument, which purports to be sufficient for the acceptance of the principles, does not seem to require that the principles be mutually acknowledged. The critical steps in the argument that require the principles be justifiable—steps (1.3) through (1.6), and step (1.10)—all seem to impose a requirement much weaker than mutual acceptability, for they seem to require only that the principles be acceptable to persons singly.

This reading of the argument is a mistake, but a subtle and instructive one. The mistake arises from too weak a reading of (1.9). To see the problem, recall that (1.8) imposes a necessary condition on principles being justifiable to each: that the principles would be chosen in a situation in which our nature determines our choice. (1.9) says that the OP is such a choice situation. (1.9) might be taken to say that the conditions of the OP are sufficient to satisfy the condition imposed by (1.8). But I think it is stronger than that, for I do not believe the Rawls of *TJ* thinks that the OP is just one of many choice situations in which our nature determines our choice. Rather, when he called the OP the "philosophically most favored" choice situation, I believe that part of what he had in mind—in *TJ*—was that *only* a choice situation that incorporates the conditions of the OP in some way is such that our nature determines our choice. Among those conditions is the publicity condition. So only a choice situation which includes that condition is one that satisfies the requirement imposed by (1.8).[7]

7. Even though Cohen's argument shows the dispensability of the OP, it appeals to what the occupants of each social position would know about the principles and their grounds; see "Democratic Equality," pp. 739 and 743. Whether our nature determines the choice of principles in Cohen's argument is a question I take up in §VII.9.

Rawls is quite clear that part of the point of the publicity condition is to have parties in the OP evaluate conceptions of justice "as publicly acknowledged and fully effective moral constitutions of social life" (*TJ*, p. 133/115). The publicity condition therefore forces each party to the OP is to ask whether the conceptions of justice under consideration could be a mutually acceptable conception of justice in a WOS. Thus the conjunction of (1.8) and (1.9), which requires that principles of justice be acceptable in the OP, also requires that principles be mutually—and not just individually—acceptable. And so while the Pivotal Argument does not explicitly appeal to the claim that society must be a cooperative scheme, the argument does appeal to premises which taken together require that the principles be mutually acknowledged.

This response to the second worry about the Pivotal Argument may be surprising. The conditions of the OP in virtue of which it satisfies the requirement imposed by (1.8)—the conditions in virtue of which it is a choice situation in which our nature is the determining element—are generally thought to include the rationality of the parties, the framing of the good by the right, and the veil of ignorance. These, it is generally thought, are the elements of the OP that represent our nature. But however much work publicity may do in other connections, it and the rest of what Rawls calls "the formal constraints on the concept of right" are not generally thought to do much of the work of representing the nature of persons asserted in (1.1). Indeed, some of the formal constraints are thought not to do much interesting work at all. For reasons that I shall explain in §VII.5, I think this is a serious mistake, one to which the objection now under consideration enables me to call attention. It is very important that the features of the OP in virtue of which Rawls endorses (1.9) include the formal constraints.

My own view is that the Pivotal Argument is wrong at one point at which it may seem unquestionably to be right and to sustain Dworkin's interpretation: between (1.5) and (1.6). The move from the former to the latter depends upon the imperative to show respect for persons. This is an imperative on which the Rawls of *TJ* is often said to rely. Reading Rawls this way is essential to Dworkin's description of Rawls's view as "rights-based." As we shall see, some philosophers think Rawls continued to rely on the imperative even after making the transition to *PL*. I shall argue in the Conclusion that political liberalism does not rely on this imperative. But it was also a mistake to suppose that Rawls relied on it in *TJ*. His never was a rights-based view, contrary to what is supposed by some of those readers who would attribute the Pivotal Argument to him. Rather, as I shall explain later, Rawls's is what he calls a "conception-based view."[8] By this he means that members of the WOS can live up to a certain conception of themselves—a conception of themselves as free and equal—only if they regulate their collective lives by mutually acceptable principles. We shall see in Chapter III that Rawls thinks members of the WOS

8. Rawls, "Political not Metaphysical," *Collected Papers*, pp. 400–401, note 19.

normally have that self-conception and want to live up to it. And so what licenses the move from (1.5) to (1.6) is not – as Dworkin would have it—a moral requirement or a right, but a claim about what they must do if they are to live up to their view of themselves.

Modulo this objection, I grant the *Public Basis View* its claim that the Pivotal Argument expresses *TJ*'s defense of the principles of justice. Since the *View* also claims that Rawls became dissatisfied with the defense of the principles he had provided in *TJ*, it implies that he became dissatisfied with the Pivotal Argument. The problem with the Argument is said to be that it is too controversial to serve as the public basis of the principles in a pluralistic society. With a clear statement of the Argument in hand, I can now state the *Public Basis View* more precisely than I did in §I.1 by pinpointing the sources of controvery and by saying exactly how the *View* claims Rawls responded to his dissatisfaction.

§I.4: The *Public Basis View* Restated

Recall that according to the first step in the Pivotal Argument

(1.1) We are by nature free and equal rational agents who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

If (1.1) expresses a metaphysical claim about the nature of persons—more specifically, a Kantian conception of the person that was supposed to compete with Aristotelian, Thomistic, Cartesian, Leibnizian, or postmodern conceptions—then it seems likely that the Pivotal Argument would prove a controversial defense of the principles of justice. For it seems likely that some members of the WOS, like some members of our own society, would be suspicious of any talk of a human nature or essence at all. Others would attack the implication that human beings are by nature prior to the ends they choose, on grounds ranging from the communitarian to the theistic. Still others would claim that human beings are naturally political and naturally participants in a common good, and would maintain that contractualist talk of individuals is an illegitimate abstraction. All these members of the WOS would object to (1.1).

Moreover, if the Pivotal Argument does begin with a Kantian conception of the person, then we would expect it to justify Rawls's principles appealing to Kantian considerations, such as the requirement to respect persons as ends in themselves and the value of autonomy. The Pivotal Argument seems to do just that for as we saw, it moves from (1.5) to (1.6) via a requirement of respect. And it moves from (1.6) via (1.7) to:

(1.8) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal persons is the decisive determining element of the choice.

(1.8) seems to say or imply that the basic structure under which we live must be governed by principles we would give ourselves. So the movement from (1.5) to (1.8) seems to show that the respect referred to at (1.5), and enjoined between (1.5) and (1.6), requires that the distributive principles under which we live be self-imposed. What really justifies C_1, and hence Rawls's principles, is that the principles satisfy this autonomy requirement. And if the real reason basic distributive principles must be self-imposed—if what, as it were, really drives the movement from (1.5) to (1.8)—is that we are and must be treated as ends in ourselves, then the requirement that we be respected seems really to be the Kantian requirement that we be treated as ends and the Kantian thought that we can be treated as ends only if we are autonomous.

Once the Pivotal Argument is read this way, we can understand why the argument ascribes only derivative force to the OP. Because the OP is a choice situation in which our nature as described in (1.1) determines our choice, as (1.9) says, the OP makes it possible to identify principles we would give ourselves. The requirement that principles be chosen in the OP—expressed by (1.10)—simply shows how to satisfy the requirements that, in matters of distribution, we be respected as ends and we give ourselves the laws under which we live. Any justificatory force imparted by choice in the OP derives from the fact that it enforces those requirements.

But if some members of the WOS would object to the Kantian expression of the person expressed in (1.1), they would also, presumably object to the Pivotal Argument's reliance on Kantian notions of respect and autonomy. So if the public justification of the principles of justice depends upon the Pivotal Argument, if (1.1) expresses a metaphysical claim, and if Kantian ethical notions and requirements are appealed to in later steps, then—from the points of view of these members of the WOS—Rawls lacks a sound public defense of the principles. The principles and their defense would not, therefore, be the objects of consensus in the WOS.

According to the stronger version of the *Public Basis View*, (1.1) *does* express a metaphysical conception of the person, the movement from (1.5) to (1.8) and beyond *does* depend upon controversial claims about respect and autonomy, and some of Rawls's critics made him realize just how controversial the Pivotal Argument would be as a result. According to the weaker version, the work of these critics made Rawls realize that his defense of the principles could be taken as relying on these controversial conceptions and claims.

To remedy his reliance on controversial premises, or to make clear that he was not relying on them, Rawls recast (1.1) as:

(1.1') We are free and equal *citizens* who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

Elsewhere in the Pivotal Argument, it is said, Rawls substituted "citizens" for "persons," so as to yield an argument that appealed to:

(1.5') Our society respects us as the kind of *citizens* (1.1') says we are only if the principles governing the ways the basic structure of our society distributes primary goods are acceptable to us as such citizens.

(1.6') Principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal *citizens*.

(1.8') The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal *citizens* is the decisive determining element of the choice.

and

(1.9') The OP is a choice situation in which our nature as *citizens* is the decisive determining element.

The result is a new public argument for C_1 and hence for the principles of justice, an argument that relies on weaker premises than the Pivotal Argument does. Rawls drew these premises from the public political culture of democratic societies. This is why the view can be presented as "free standing," as standing free of metaphysical claims about the nature of persons. Once (1.1) and (1.5) were weakened, the crucial transition to the sixth step of the argument could be weakened as well. In the original version of the argument, the move from (1.5) to (1.8) seemed to be driven by the value of personal autonomy. In the revised argument, the move from (1.5') to (1.8') can appeal to autonomy in political life. Rawls was then able to argue that reasonable people in a democratic society, even with different views of the good, could accept the weakened premises and inferences. The revised version of the Pivotal Argument could therefore serve as the shared public defense of the principles of justice in a WOS. According to the strong version of the *Public Basis View*, the premises of the public argument for C_1 were introduced as part of a modification of justice as fairness. According to the weaker version of the *View*, they were introduced to clarify and to remedy misunderstandings.

The *Public Basis View* may now seem quite appealing. For one thing, the changes in the Pivotal Argument that the *View* identifies do seem to be reflected in Rawls's texts. For example, the shift in starting points from (1.1), which is a claim about our nature as persons, to (1.1'), which is a claim about our nature as citizens, seems to be reflected in the later Rawls's insistence that justice as fairness begins from a political conception of the person, which he equates with a conception of the citizen.[9] The claim that the move from the fifth to the eighth step in the argument depends upon a weaker form of autonomy than personal autonomy fits with Rawls's insistence in *PL* that the autonomy enjoyed by members of the WOS is "political not ethical" (*PL*, p. 77). The shift

9. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 397.

from (1.9), according to which the OP brings our nature as persons to bear on the choice of principles, to (1.9'), according to which it brings our nature as citizens to bear on that choice, seems to be reflected in *PL*'s insistence that the OP is a "device of representation" "in which *citizens*' moral powers . . . are modeled" (*PL*, p. 48, emphasis added). So the *Public Basis View* enjoys some textual support. The *View* also explains why Rawls made the changes these shifts reflect. Furthermore, it stands to reason that the changes between *TJ* and *PL* would alter the argument Rawls offered for the principles of justice. The *Public Basis View* shows precisely how they would.

Rawls is said to have recast the public defense of his principles so that it could be the object of consensus in a pluralistic society. The Rawls of *PL* seems to say that that consensus would be achieved by "overlap." The *Public Basis View* goes hand in hand with a compelling picture of consensus achieved in this way. According to both versions of the *Public Basis View*, the picture of an overlapping consensus is one of different comprehensive views providing, as it were, deep arguments—sometimes deductive and sometimes not—for the weakened premises of the reformulated argument for C_1. This picture—and hence the *View*, which suggests it—enjoys some textual support. It is suggested, for example, by one of Rawls's most memorable and graphic descriptions of an overlapping consensus, in which he likens an overlapping consensus to a set of theorems implied by different axioms.[10] The Rawls of *PL* implies that the idea of an overlapping consensus is introduced to explain the stability of a WOS (*PL*, p. 141). The *Public Basis View*'s picture of an overlapping consensus can, if pressed, yield a view about how the existence of an overlapping consensus contributes to stability. For it would be natural to conclude from the picture that such a consensus contributes to stability because when an overlapping consensus obtains, citizens have moral sources of deeper and more stable conviction for the weakened premises of the Pivotal Argument, and hence for the principles themselves. By suggesting the picture, the *Public Basis View* seems to possess an explanatory power that tells in its favor and that fits with Rawls's purposes in introducing the idea of an overlapping consensus.

I believe that many readers of Rawls accept the strong—or more often—the weak—variant of the *Public Basis View* as the best explanation of the differences between *TJ* and *PL*. I shall not try systematically to locate the *Public Basis View* in the voluminous literature on Rawls, though I believe the *View* has some very prominent defenders.[11] Instead, I shall largely rely on the

10. See Rawls, "Political Not Metaphysical," *Collected Papers*, p. 411; John Rawls, "The Idea of an Overlapping Consensus," *Collected Papers*, pp. 521–48, p. 430.

11. I believe the explanations of Rawls's political turn offered by Charles Larmore and Bruce Ackerman are—when fully spelled out—best interpreted as versions of the *Public Basis View*; indeed, if I have read Larmore correctly, the *View* gets its most sophisticated expression in his hands. See Charles Larmore, "Political Liberalism," *Political Theory* 18 (1990): pp. 339–60, pp. 345–46 and Bruce Ackerman, "Political Liberalisms," *The Journal of Philosophy* 91 (1994): pp. 364–86, p. 365.

reader's sense that the *Public Basis View*, particularly its weaker variant, is the prevailing interpretation.

Despite the appeal of the *Public Basis View* and the prominence of the interpreters who suggest some of its essentials, there are serious textual and philosophical difficulties with the *View* in both its strong and its weak variants. I shall expose those difficulties in the next two sections. But before I do so, let me be clear about where the problems with the *View* lie. As I said at the end of §1.3, they do not lie in the imputation of the Pivotal Argument to the Rawls of *TJ*. Nor do they lie in the claim that the Rawls of *PL* weakened the public argument for the principles in the ways described in this section, so that it relies on (1.1') and (1.9'). Nor, finally, do they lie in the *View*'s implication that when an overlapping consensus obtains, members of the WOS may endorse the crucial steps of the Pivotal Argument for deeper reasons of their own. Rather, the most serious difficulties with the *Public Basis View* lie in what it says about why Rawls took his political turn and about the way an overlapping consensus stabilizes a WOS.

On my reading, Rawls made the changes between *TJ* and *PL*, not because he was dissatisfied with the Pivotal Argument or with the possibility that that Argument lent itself to misreading, but because he became dissatisfied with his treatment of stability in part III of *TJ*. It was his attempt to remedy the problems with that treatment—rather than any problems with the Pivotal Argument—that led him to shift from (1.1) and (1.9) to (1.1') and (1.9'). And according to my reading, an overlapping consensus stabilizes, not just by bringing about the acceptance of those premises and other claims in the Pivotal Argument, but also by removing certain temptations to defect from the agreement reached in the OP. I shall begin to defend this reading in Chapter II. Now I want to show the difficulties with the *Public Basis View*, beginning with the strong version.

§I.5: Difficulties with the Strong Version

According to the strong version of the *Public Basis View*, (1.1) expresses a *metaphysical* conception of the person. By now, Rawls's own later interpretations of his own work have brought about widespread consensus that it does not. But even those who subscribe to this consensus are not always clear about exactly how and why Rawls argues for a negative answer. I want to look at that argument, not only to show what is wrong with the strong version of the *View*, but also to lay some groundwork for what I shall say about (1.1) when I lay out my own reading of Rawls.

Rawls's argument depends upon a basic distinction that he borrowed from H. L. A. Hart, the distinction between a *concept* and its various *conceptions*. Just as we have a concept of justice which can be specified into various conceptions (*TJ*, p. 5/5), so—Rawls thinks—we have a concept of the person which can be specified into various conceptions. The concept of the person is

specified into a conception by giving an account of the powers, interests and properties persons have as such, or the standards by which human beings and actions are assessed. To specify a metaphysical conception of the person is to specify the concept of the person by giving an account that draws on theses and principles from metaphysics, or on the answers to metaphysical questions. To specify an ethical conception of the person is to specify the concept by giving an account that draws on values and theses from moral philosophy. To specify a political conception of the person is to specify the concept by giving an account that draws on political values and theses from what Rawls came to call "the domain of the political."

It may seem difficult to say just what a metaphysical conception of the person is, on this understanding, because it seems difficult to say exactly what a metaphysical thesis about persons is and how it is to be distinguished from an ethical thesis. I believe Rawls relies on disciplinary boundaries to draw the needed distinctions. The discipline of metaphysics, he thinks, concerns itself with a set of questions about persons—for example, about their identity across possible worlds or their continuity through time—that can be distinguished from the questions that are taken up by other subdisciplines within philosophy. In denying that he relied on a metaphysical conception of persons, Rawls does not mean to deny that he relied on (1.1) or that (1.1) expresses a conception of the person properly so called. He means, rather that (1.1) does not express a conception that is specified by drawing on theses from the discipline of metaphysics. That is why he says that the conception of the person on which justice as fairness relies "is not taken from metaphysics or the philosophy of mind, or from psychology; it may have little relation to the conceptions of the self *discussed in those disciplines*."[12]

Why rely on disciplinary boundaries to distinguish metaphysics from other areas of philosophy? The answer, I think, lies in the real point Rawls is trying to make by denying that he relies on a metaphysical conception of the person. That point concerns the independence of moral and political philosophy from certain clearly identifiable problems and questions that are now thought to fall within the domain of metaphysics and the philosophy of mind. Rawls thinks that progress in political philosophy need not be hostage to the outcome of debates about personal identity, for example, because political philosophy can specify a conception of the person—such as (1.1)—without assuming the answers defended by one or another party to those debates.[13] Rawls expresses this point in his later essays with the vague denial that he is relying on a metaphysical conception of the person. To understand what he means, we have to bear in mind the point I have said he really wants to make. In light of that point, the disciplinary view of metaphysics he works with suffices.

12. John Rawls, *Justice as Fairness: A Restatement* (Cambridge, MA: Harvard University Press, 2001), ed. Erin Kelly, p. 19.

13. See John Rawls, "The Independence of Moral Theory," *Collected Papers*, pp. 286–302, especially pp. 295–301.

Of course, one could adopt a different and broader view of metaphysics, and argue that Rawls relies on metaphysical assumptions after all.[14] This conclusion might not be wrong, but it would miss the point Rawls is trying to make by denying that (1.1) expresses a metaphysical conception. Once that point is clear, and it is clear that the point depends upon a quite specific understanding of metaphysics, it is also clear that the categories "political" and "metaphysical" do not exhaust the kinds of conception there can be, since there are other sub-disciplines in philosophy besides metaphysics and political philosophy. Rawls's denial leaves open the possibility that (1.1) expresses an *ethical* conception of the person, one which specifies our concept of the person by drawing on some theory in moral philosophy such as Kantianism or by appealing to some ethical value such as autonomy.

For those who want to explain the changes between *TJ* and *PL*, the question of whether justice as fairness relies on an ethical conception of the person is far more interesting than the question of whether (1.1) is metaphysical. I shall return to this question in Chapter III. For now, suffice it to say that I agree with a large number of commentators in thinking that nothing forces us to read the Rawls of *TJ* as relying on a metaphysical claim about the nature of persons. We do not have to read the remarks in *PL* to the effect that the fundamental ideas of justice as fairness are drawn from the public culture of democratic societies as repudiations of metaphysical claims—such as (1.1) is taken to be—on which Rawls previously relied to establish the principles. The strong variant of the *Public Basis View* is right to suppose that the Rawls of *TJ* accepted (1.1), but it goes wrong by misinterpreting that claim.

The strong version of the *Public Basis View* also holds that in *TJ*, Rawls moved from the Kantian conception of the person expressed in (1.1), via the requirement to respect persons as ends, to

(1.6) The principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal persons.

The move from this claim to:

(1.10) The principles governing the ways the basic structure distributes primary goods must be acceptable in the OP.

is then said to have been driven by the controversial value of autonomy—a move Rawls had to revise, by introduction of the weaker notion of political autonomy, once he realized how controversial the original appeal to Kantian autonomy would be.

As we saw, this reading depends upon the fact that the Pivotal Argument moves to (1.10) from:

14. In her careful defense of Rawls against communitarian critiques, Amy Gutmann says that "Rawls must admit this much metaphysics—we are not radically situated selves." See her "The Communitarian Critique of Liberalism," *Philosophy and Public Affairs* 14 (1985): pp. 308–22, p. 314.

(1.8) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal persons is the decisive determining element of the choice.

and the claim, expressed in (1.9), that the OP is a choice situation in which our nature is the decisive determining element. It is these two claims that show why principles adopted in the OP are principles we would give ourselves.

The problem with this reading is that in *TJ*, (1.10) is overdetermined. Rawls justifies the requirement that principles be adopted in the OP in a number of ways. Some of the arguments for that requirement depend upon what Rawls calls the "Kantian Interpretation" of justice as fairness, laid out in *TJ*, §40. The argument for (1.10) that goes by way of (1.8) and (1.9) is one such argument. But Rawls is careful to distinguish those arguments for (1.10) from other arguments for it (see *TJ*, pp. 139ff./120ff). Some of those arguments appeal to our intuition that arguments for principles of justice should not appeal to considerations that are irrelevant from a moral point of view (*TJ*, p. 141/122), and that the OP draws together "the restrictions that it seems reasonable to impose on arguments for principles of justice" (*TJ*, p. 18/16). Our intuitions about such restrictions are, I believe, supposed to have sufficient force to justify the requirement that principles be acceptable in the OP, independent of the value of autonomy. If this is right, then the Rawls of *TJ* thought it possible to bypass (1.8) and (1.9) and proceeds directly to (1.10) from:

(1.7) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation that is uninfluenced by natural and social contingencies.

And if that is right, then it is a mistake to think that in the WOS of *TJ*, the public defense of the principles of justice *required* appeal to (1.8), (1.9), or appeal to the value of autonomy. But this thought is one of the central tenets of the strong version of the *Public Basis View*.

One reason I included (1.8) and (1.9) in the Pivotal Argument is that doing so enabled me to make the *Public Basis View* precise and so to show why readers might be drawn to it. Another is that I assume the public culture of the WOS would make the Kantian Interpretation of justice as fairness and Kantian arguments for the OP available. Incorporating (1.8) and (1.9) into the Pivotal Argument was an economical way of showing one of the ways the public culture might do that. The importance of showing that brings me to the third reason for including them. (1.9) does indeed imply that principles chosen in the OP are principles we would give ourselves. It therefore shows why Rawls thinks members of the WOS would act autonomously in acting *from*, and not merely *in accordance with*, the principles. The Rawls of *TJ* did not treat the fact that they would as an indispensable step in the public defense of the principles. But he did—as we shall see—think public knowledge of that fact was essential to showing that members of the WOS would judge that acting from

the principles of justice is part of their good. That fact is made publicly available in the WOS of *TJ* when (1.9) is included in the publicly available justification of the principles.

§I.6: Difficulties with the Weak Version

Proponents of the weak version of the *Public Basis View* assume what objections to the strong version are trying to show—that (1.1) was not meant to express a metaphysical conception of the person. According to the weak version, the differences between *TJ* and *PL* result from Rawls's attempt to make that point clear. So the objections to the strong version that I canvassed in the previous section do not tell against the weak version. Indeed, they may seem to lend it more plausibility, since—if they are successful—they establish one of the fundamental assumptions of the weak version. But there are serious problems with that version as well. Some of those problems arise from trying to square the weak version of the *Public Basis View* with Rawls's texts.

The weak variant of the *Public Basis View*—like the strong one—claims that the presentation of justice as fairness as a political liberalism was a response to communitarian criticisms or misreadings. But Rawls explicitly denies that the move from *TJ* to *PL* was motivated by communitarian critiques of *TJ* (see *PL*, p. xix, note 6). Rather, he says that he first began thinking about revising the view laid out in *TJ* when he read the draft of an article by Samuel Scheffler (*PL*, pp. xxxiv–xxxv). Scheffler's is a very short piece to have motivated such significant changes in Rawls's view. The reasons Rawls found this piece so provocative have remained somewhat obscure. The line of thought that Rawls followed from Scheffler's article to *PL* is difficult to trace. It can be discerned, if it all, only after the explanation of Rawls's transition is already in place.[15] But however that line is plotted, it most definitely does not go by way of a communitarian critique.

The communitarian misreadings that Rawls wanted to discredit are, according to the weak version of the *Public Basis View*, misreadings of *TJ*, part I, where – as we saw—critics like Michael Sandel noted that Rawls relied on (1.1) to set up the OP. Thus according to the weak variant, the differences between *TJ* and *PL* are introduced to clarify that part of *TJ*. But this ignores what Rawls himself says about why he made the transition to *PL*. Speaking of the essays in *PL*, Rawls says:

> Indeed, it may seem that the aim and content of these lectures mark a major change from those of *Theory*. Certainly, as I have indicated, there are important differences. But to understand the nature and extent of the differences, one must see them as arising from trying to resolve a

15. See §VIII.4.

> 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. (*PL*, pp. xv–xvi)

This remark directs us, not to the considerations used to justify the original position in part I of *TJ*, but to a very different part of the book that is much less frequently read: *TJ*, part III.

Since the weak version of the *View* implies that Rawls made the changes between *TJ* and *PL* simply to restate and clarify what he meant in *TJ*, it seems to be committed to the claim that Rawls's treatment of stability was not really changed with his political turn. The remark I have just quoted seems to contradict this claim. Furthermore, if the weak version were correct to imply Rawls's treatment of stability did not change—and if the picture of stability suggested by the weak version is correct—then Rawls must always have meant that the stability of a WOS is achieved by an overlapping consensus on the shallow or weakened premises of the Pivotal Argument. He must have introduced the idea of an overlapping consensus simply to make explicit that this account of stability is what he had in mind all along. But this is something Rawls explicitly denies. In a crucial passage at the end of "Political Not Metaphysical," Rawls says of his account of stability in *TJ* that "the account . . . was not extended, as it should have been, to the important case of an overlapping consensus, as sketched in the text; instead this account was limited to the simplest case."[16] The idea of an overlapping consensus was an innovation, not—as the weak version of the *Public Basis View* seems to imply—a clarification.

The weak version of the *View* faces philosophical as well as textual difficulties. One is that its explanation of the changes between *TJ* and *PL* is too simplistic. To see this, note first that the *Public Basis View* introduces several crucial claims as assumptions that are, at best, examined lightly. Those claims include (i) that metaphysical claims about persons and appeals to autonomy would be the objects of reasonable disagreement in the WOS, and (ii) that premises which are the objects of reasonable disagreement cannot serve as the public basis for a conception of justice. The proponent of the weak variant thinks that (ii) is obvious, and that (i) follows directly from another claim she simply assumes, (iii) the "fact of pluralism." Helping herself to (ii) and to the inference from (iii) to (i), the proponent of the weak variant of the *Public Basis View* claims she can explain the changes between *TJ* and *PL*.

But consider the support the proponent of the weak variant offers for (i). Rawls undoubtedly accepts (iii), the fact of pluralism. But he does not do what the proponent of the *Public Basis View* does, which is simply to assume it. The fact of pluralism, as Rawls came to understand it, is the fact of *reasonable* pluralism. According to the fact of reasonable pluralism, the diversity of

16. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 414, note 33.

reasonable views is "in part the work of free practical reason within the framework of free institutions" (*PL*, p. 37). But what is it about free institutions that gives rise to pluralism? Is it just that free institutions do not enforce any one religion, say? Or is there something more to be said about the connection?

As I shall indicate at greater length later, these are questions about which Rawls thought deeply. Rawls believed that under the free institutions—including the free institutions of a WOS—citizens would have a certain view of themselves that those institutions encouraged: they would think of themselves as, in various ways, free. It is in part because free institutions encourage members of a WOS to think of themselves as free that they exercise their practical reason as they do and that the fact of reasonable pluralism obtains. This important causal connection between the self-conception encouraged by the institutions of a WOS and the fact of reasonable pluralism is overlooked when the fact of pluralism is simply assumed to obtain.

Consider now whether (ii) is as obvious as it might seem to be. I believe Rawls regarded (ii), like (i), as the conclusion of an argument, one premise of which concerns the view citizens of a WOS would have of themselves. (ii) is true, he would maintain, because the citizens of a WOS are encouraged by their institutions to think of themselves and one another as free equals who are, as such, worthy of being offered reasons they can accept. This point is likely to be overlooked when (ii) is assumed uncritically, as if it needed no argument.

How did Rawls think the basic institutions of the WOS encourage members of that society to think of themselves as I have said they would—as free equals? Rawls recognized in *TJ*, and said even more clearly in the original *Dewey Lectures*, that the public justification of the principles—and especially the conceptions and ideals of the person that are part of that justification—has an important influence on how members of the WOS think of themselves. It is because the institutions of liberal democracies treat citizens according to principles suitable for a society of free equals, and justify their treatment of them as such, that members of liberal democracies think of themselves as free—and, in the WOS, as autonomous. And so Rawls thought that the conceptions and ideals of the person that are part of public justification play an important role in bringing about the fact of reasonable pluralism,[17] and the fact that citizens of a

17. My claim that pluralism results from the way a WOS encourages citizens to think of themselves may seem to contradict the explanation of pluralism that Rawls himself offers. In *Restatement*, for example, he asks "how might reasonable disagreement come about?." He answers by listing what he calls "the burdens of judgment." I do not think my answer contradicts Rawls's, since I think Rawls must be understood as asking how reasonable pluralism might come about *among persons who regard themselves as free*. This reading gets some confirmation from Rawls's discussion of the burdens of judgment in *PL*. There Rawls says that reasonable pluralism arises among reasonable persons, and reasonable persons are free and equal persons (see *PL*, p. 55).

WOS expect to be offered acceptable justifications for their basic political arrangements.

Once we see that the *Public Basis View* overlooks this formative role and its implications for pluralism and public justification, we can see that the *View* is bound to offer too simplistic an account of why Rawls would have wanted to disavow the Pivotal Argument, and therefore too simplistic an account of the changes between *TJ* and *PL*. For according to the *View*, Rawls thought that a metaphysical conception of the person like that said to be asserted in (1.1), and claims about autonomy like those connected with (1.8) and (1.9), could not provide the public basis of justification simply because they would not be accepted in a pluralistic society like the WOS. He then based justice as fairness on a political conception of the person because it provided a shared basis. But though Rawls did indeed worry that (1.1) and (1.9) would be controversial in a pluralistic society, his worries about them ran much deeper than the weak version of the *Public Basis View* alleges. He worried that (1.1) and (1.9) were too strong because of the long-term effects of institutionalizing principles publicly based on them. Institutionalizing those principles and publicizing their justification would themselves, he thought, encourage the pluralism that resulted in disagreement about (1.1) and (1.9), and would themselves encourage citizens' sense that they are owed a justification that is not so controversial. These consequences could ultimately lead to disagreement about C_1.

Thus Rawls's worry about the Pivotal Argument was not simply that it rests on a basis which would not be the object of consensus as the *Public Basic Views* alleges. His worry was that using that argument as the public defense of the principles of justice would itself undermine support for them. And so his worry about the Pivotal Argument was not just that it has controversial premises, but that the effect of publicizing it would be to destabilize the justice of the WOS. This concern bulks large for Rawls. As we shall see in §II.1, one of Rawls's aims both early and late was to show that valid principles of justice, when institutionalized and publicized, stabilize themselves.

One reason that the *Public Basis View* offers too simplistic an explanation of the changes between *TJ* and *PL* is that it operates with too superficial an understanding of the publicity condition. When a conception of justice satisfies that condition, one effect is that citizens can come to know the bases on which the conception is supposed to be publicly accepted. This effect is what the proponent of the *Public Basis View* draws on when he says why he thinks Rawls wants to disavow the Pivotal Argument. But there are other consequences of publicity as well, consequences that play a role in *TJ* but that Rawls is much clearer about it in the original *Dewey Lectures* and beyond. As I have already suggested, when a conception of justice is public, it has an educational or formative role. The publicity of its conceptions and ideals of the person or of the citizen encourages citizens of the WOS to think of themselves in that way, or to aspire to be that kind of person. This is part of how justice as fairness,

when implemented by public institutions, generates support for itself. Exactly how this happens, and what is consequences are, are topics I shall take up in §III.3 and in Chapter VIII. What matters for present purposes is that the proponent of the *Public Basis View* cannot make good on his own explanation of Rawls's dissatisfaction with the Pivotal Argument without appreciating this consequence of publicity. And as we shall see, we cannot understand Rawls's reasons for becoming dissatisfied with his original treatment of stability in part III of *TJ*—nor will we understand Rawls's reformulated argument for it—without appreciating it.

Why can't the changes between *TJ* and *PL* be explained by the more nuanced account I have suggested—the worry that a public argument for the principles which relies on (1.1) and (1.9) would be self-defeating? What is wrong with the claim that Rawls weakened (1.1) and (1.9) so that he could argue for C_1 from premises on which there could be an overlapping consensus? The problem with this explanation is one I believe it shares with the *Public Basis View*. That problem lies in the way both explanations suppose that an overlapping consensus helps to stabilize a conception of justice.

Earlier, when I introduced the picture of an overlapping consensus associated with the *Public Basis View*, I said it would be natural to conclude that according to the *View*, an overlapping consensus contributes to stability because when such a consensus obtains and is known to obtain, it is public knowledge that all citizens have deep, and presumably stable, belief in the premises of the argument for C_1. This is rather a cerebral take on the work that is done by an overlapping consensus. While the interpretation may be correct as far as it goes, it does not go very far. For stability depends upon citizens' having a desire to do justice that is effective, a desire that is strong enough to overcome temptations to act unjustly. According to the intellectualist interpretation of an overlapping consensus that we are now considering, when such a consensus obtains and is known to obtain, each member of the WOS believes, and believes that everyone else believes, the principles of justice are valid. But the interpretation does not say anything about how these beliefs are connected with an effective desire to be just. As we shall see, showing the connection is an important part of what the idea of an overlapping consensus was introduced to do. An interpretation of an overlapping consensus that leaves these connections out of account, or simply assumes that they obtain when everyone is convinced by the public defense of the principles, thereby omits something very important. If we are too taken with any such interpretation, we will miss much that is important about Rawls's account of stability. We will therefore—I shall contend—miss much that is important about why Rawls made the changes between *TJ* and *PL*.

§I.7: Conclusion

Despite its appeal and popularity, the *Public Basis View* faces a number of daunting obstacles. In its strong version, it requires that we impute metaphysical

claims to Rawls which his texts do not force upon us. In its weaker variant, it is inconsistent with what Rawls says about why the idea of an overlapping consensus was introduced and about why he made the changes he did between *TJ* and *PL*. It does not say enough about why Rawls would modify the Pivotal Argument because it rests upon too superficial an understanding of the publicity condition. And it suggests too cerebral an account of political stability. I believe that these problems, when taken together, are insurmountable.

I have granted that the Pivotal Argument would be part of the public culture of the WOS of *TJ*. In his later work, Rawls clarified and weakened some of its crucial claims, and so I grant that the public culture of the WOS of *PL* would contain a "political version" of the Pivotal Argument. But the thesis that dissatisfaction with the Pivotal Argument is what motivated the transition from *TJ* to *PL* cannot withstand scrutiny. Definitive refutation of the *Public Basis View* depends, however, upon providing and defending some other explanation for the changes between *TJ* and *PL*. That is what I propose to do in the chapters that follow.

II

Stability and Congruence

To understand the changes Rawls made between *TJ* and *PL*, we need to identify an argument or set of arguments with which he became dissatisfied or from which he wanted to distance himself, and to say what he found unsatisfactory about them. In Chapter I, we saw that the *Public Basis View* is not up to these tasks. In its strong variant, the *View* explains the changes Rawls introduced by claiming that he first relied upon and then repudiated metaphysical premises of what I called the Pivotal Argument. The premises in question were not, however, ever intended as metaphysical claims. According to the weak variant of the *View*, Rawls introduced the changes between *TJ* and *PL* to make clear that he had not ever relied upon the premises, so understood. The proponent of the weak variant thus grants that the strong variant is mistaken, but maintains that Rawls recast justice as fairness as a political liberalism in order to correct the misreading of part I of *TJ* on which the strong variant depends. But the weak variant of the *Public Basis View* also faces a number of difficulties. As we saw, it gives too simplistic an account of why Rawls would have wanted to dissociate himself from the Pivotal Argument, metaphysically interpreted. It also lends itself to too cerebral an account of stability. But the most obvious of the difficulties faced by the weak variant of the *Public Basis View* is textual. It is at odds with what Rawls himself says about the reasons for the changes he made in his view.

Recall that in describing why he made the changes between *TJ* and *PL*, Rawls directs our attention to part III of *TJ*. He says that he made those changes

> to resolve a serious problem internal to justice as fairness, namely . . . the fact that the account of stability in part III of *Theory* is not consistent with the view as a whole. (*PL*, pp. xv–xvi)

I said in Chapter I that I think the *Public Basis View* offers the most widely accepted explanation of the changes between *TJ* and *PL*, but I do not mean to suggest that it is the only explanation. Some readers have taken Rawls at his word about why those changes were made and have tried to locate the reasons for those changes in the part of *TJ* to which Rawls directs us. According to one interpretation, Rawls came to see that the principles of justice would encourage freedom of thought, and hence pluralism, and that if the well-ordered society (WOS) were pluralistic, then some reasonable members of the WOS would disagree with the principles. This, it is said, is inconsistent with Rawls's early treatment of stability because according to the treatment of stability in *TJ*, everyone in the WOS accepts the same principles of justice.[1] According to another interpretation, Rawls came to see that one of his critical arguments in part III could not succeed if the WOS were pluralistic, and it was the failure of that argument that undermined the treatment of stability.[2]

For reasons I shall give much later, I do not find these interpretations plausible. My view is that Rawls made the changes between *TJ* and *PL* because he came to see that a major part of his treatment of stability—namely, his treatment of what he calls "congruence"—was unsuccessful. Not only did he come to think that the arguments he offered for congruence failed, but he came to think that the problem of congruence as he posed it in *TJ* was misconceived. This chapter lays some of the groundwork for my reading. Since Rawls says that he made the turn to political liberalism because of problems in *TJ*'s treatment of stability, I begin by asking what the discussion of stability was supposed to show. As we shall see, Rawls's attempt to show that justice as fairness would be stable is extraordinarily ambitious.

§II.1: Stability, Inherent and Imposed

What Rawls said about the problem of stability is easily misunderstood. In the preface to *PL*, Rawls says of his attempt to correct *TJ*'s account of stability:

> Surprisingly, this change in turn forces many other changes and calls for a family of ideas not needed before. I say surprisingly because 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 force such extensive revisions. (*PL*, p. xix)

1. Burton Dreben, "On Rawls and Political Liberalism," in the *Cambridge Companion to Rawls* (Cambridge: Cambridge University Press, 2003), ed. Samuel Freeman, pp. 316–46, p. 317.

2. Samuel Freeman writes "what is primarily unrealistic about the account in *Theory*, I conjecture, is the Kantian congruence argument." See his "Congruence and the Good of Justice," in his *Justice and the Social Contract* (New York: Oxford University Press, 2007), pp. 143–72, p. 168.

This remark is itself surprising, since philosophers since Plato have been concerned with the question of how to maintain order within political society. Within the contract tradition, Hobbes seems obviously to be concerned with how social stability is to be maintained.[3] To see what Rawls had in mind in this passage, we need some distinctions.

The kind of stability that concerned Rawls is different from stability as it is commonly understood in political science literature. What is usually discussed there is what we might call *state stability*. Let us say that this kind of stability obtains in a state S for some period just in case there is no significant extra-constitutional change in S's borders or in the structure of its government in that period, and there is regular compliance with the law by a sufficiently large portion of S's population. State stability can obtain even if the laws with which citizens comply are unjust. Indeed, as history shows, totalitarian states can be stable for considerable periods of time.

States that are stable can be contrasted with states or cooperative schemes that remain, if not perfectly just, then just or approximately so, over time.[4] Let us call such states *stably just*. The stability with which Rawls is concerned is this kind of stability, rather than *state stability*. This interest is confirmed by his remark that in the context of *TJ*, part III: "stability means that however institutions are changed, they still remain just or approximately so" (*TJ*, p. 458/401). When a society is just or approximately just at some time, it is—for the moment at least—effectively regulated, and publicly known to be effectively regulated, by a valid public conception of justice.[5] It is then in a condition of general equilibrium: everyone knows that everyone else acts justly, and each replies to the justice of others by being just himself.[6] But not all general equilibria are stable. A state or a scheme of cooperation is stably just when it is in a just general equilibrium that is stable, so that a valid conception of justice effectively regulates it, and is known effectively to regulate it, over time. Thus we might say that Rawls is concerned, in the first instance, with "equilibrium" and "stability" as they are predicated of conceptions of justice. When a conception of justice is in a stable equilibrium, the institutions it regulates will be stably just.

Conceptions of justice can be stabilized in at least two ways. I shall refer to the two kinds of stability that result as *inherent* and *imposed*. As we shall see, Rawls thinks that while Plato, Hobbes and many other political philosophers may

3. See Brian Barry, "John Rawls and the Search for Stability," *Ethics* 105 (1995): pp. 874–915, p. 880.

4. For the notion of perfect justice, see *TJ*, p. 78/68.

5. Here I ignore the complication, introduced in *PL*, that the WOS may be effectively regulated by a family of liberal political conceptions of justice rather than by just one.

6. More specifically, as we shall see in §X.7, it is in a state of wide and general reflective equilibrium. For a comparison between stability and general equilibrium in economic theory, see Rawls, "Independence of Moral Theory," *Collected Papers*, p. 294.

have been concerned with how societies can be stably just, they have generally thought that stability of this kind needed to be *imposed*. By contrast, in *TJ*—and, modulo some qualifications, throughout his published work—Rawls wants to show that his conception of justice, justice as fairness, would be *inherently* stable. How is the difference between inherent and imposed stability to be understood?

Crudely put, a conception of justice is inherently stable if a society that is well-ordered by it generally maintains itself in a just general equilibrium and is capable of righting itself when that equilibrium is disturbed. Rawls says that in a WOS, "inevitable deviations from justice are effectively corrected or held within tolerable bounds *by forces within the system*" (*TJ*, p. 458/401). To grasp the difference between inherent and imposed stability, then, we have to see what the boundaries are of "the system" to which Rawls refers.

Rawls famously takes what he calls the "basic structure of society" as the primary subject of justice. He says that the basic structure consists of major social institutions "taken together as one scheme" (*TJ*, p. 7/6). It would be natural to read "the system" as referring to the basic structure of society and to define inherent stability as stability that depends exclusively upon forces within the basic structure. It would then be natural to conclude that Rawls thinks a conception of justice is inherently stable when a basic structure that conforms to it relies only on the forces at its disposal to correct deviations from justice and to hold injustice "within tolerable bounds."

This interpretation, though natural, is misleading. In "Distributive Justice: Some Addenda"—a paper published just three years before *TJ*—Rawls says "a conception of justice is stable if *the institutions which satisfy it* tend to generate their own support, at least when this fact is publicly recognized."[7] This remark does not just apply to Rawls's own conception of justice, or to the society well-ordered by it. It applies to conceptions of justice and societies generally. Once we see that, we can see the problem with reading "the system" as referring to the basic structure. The basic structure includes a society's governing apparatus. Some parts of some societies' governing apparatus—such as repressive penal institutions—might be established precisely because the institutions which satisfy the conception of distributive justice in question are *incapable* of generating their own support. And so those elements of their basic structures are established to bring about stability that would be impossible without them. The stability that relies on these elements of the basic structure is therefore not inherent, but imposed. As we shall see, the basic structure of Rawls's WOS will not include elements the purpose of which is to impose stability. But because Rawls explicitly contrasts the inherent stability of the WOS with that of societies in which the stability of conceptions of justice is imposed, and because taking "the system" as "the basic structure" can blur that crucial contrast, I want to propose a different interpretation.

7. John Rawls, "Distributive Justice: Some Addenda," *Collected Papers*, pp. 154–75, p. 171 (emphasis added).

What, then, is "the system" within which stabilizing forces must originate if justice as fairness is to be inherently stable? On my reading, it consists of the basic institutions that would be established to implement the principles of distributive justice. They are the institutions discussed in *TJ*, part II. As I hinted at the end of §I.6, they also include the institutions by which justice as fairness is publicized and by which members of the WOS are educated in it. When Rawls tries to show the inherent stability of justice as fairness, what he is trying to show is that these institutions, taken together, *would* generate their own support. How Rawls thinks they would do that is the subject of much of this book. In this chapter, I shall say just enough to clarify the distinction between inherent and imposed stability, to provide an initial statement of the problem of stability as Rawls conceives it and to suggest why Rawls thought his early treatment of that problem failed.

In *TJ*, Rawls maintains that the public institutions of a WOS are effectively regulated by a conception of justice only if all the members of the WOS accept that conception, where "acceptance" entails that the members of the WOS all willingly do their part to uphold their just institutions and to restore justice when injustice occurs. One of the ways the institutions discussed in part II of *TJ* generate their own support is by fostering a sense of justice, so that the citizens who live under them are disposed to do these things. Later, I shall comment on the process of moral development Rawls sketches in *TJ*. For now, I shall assume that that process—and the role that just institutions play in it—are familiar enough. What matters for present purposes is this. Even a successful argument that the institutions of a WOS would foster a sense of justice is not enough to show that they successfully generate their own support, and so it is not enough to show that justice as fairness is inherently stable. It is not enough because members of the WOS can still decide not to act from their sense of justice. Even if everyone is shown to have an effective sense of justice, at least two threats to stability remain—as Rawls notes quite explicitly (*TJ*, pp. 336/295–96, 497f./435).

At one point in *TJ*, Rawls observes that "The sense of justice leads us to promote just schemes and to do our share in them *when we believe that others, or sufficiently many of them, will do theirs*" (*TJ*, p. 267/236). Thus, the sense of justice is founded on reciprocity. If I believe that others will act justly—by paying their taxes, for example—then my sense of justice will incline me to do my share as well. But stability is threatened if citizens lack sufficient reason to think others will do their share, for then their sense of justice may not have that effect. So the WOS faces what I shall call the *mutual assurance problem*[8]: if

8. This problem would normally be called "the assurance problem." I have chosen a different label because in some of the game-theoretic literature to which Rawls refers, the label is used to designate games in which the second threat to stability has already been overcome. See Amartya K. Sen, "Isolation, Assurance and the Social Rate of Discount," *The Quarterly Journal of Economics* 81, 1 (1967): pp. 112–124; for Rawls's reference to this article, see *TJ*, p. 269, note 8/237, note 7.

citizens are to act from their sense of justice consistently, each must have some assurance that others will consistently act justly as well. If the first threat is to be averted, the WOS must provide that assurance.

To see the second threat, imagine that I know everyone else will act justly. Then I may be tempted to "free-ride" on their justice by behaving unjustly myself. For example, if I know that everyone else will pay their taxes, I may be tempted to cheat on mine, confident that other people's compliance guarantees the government has sufficient revenues. Cheating would require me to ignore the promptings of my sense of justice, but the prospect of extra money may prove so attractive that I am willing to do so. Of course, no one person's cheating will undermine the justice of the WOS and render its conception of justice unstable. But if everyone, or large numbers of people, reason similarly, then they will not pay their taxes either. In that case, the conception of justice *will* be destabilized.

It is important to see how this threat to stability comes about. The principles of justice are chosen in the OP. The principles for institutions and the tax laws—all designed to implement the principles—are adopted in a "four-stage" sequence of rational choice situations (*TJ*, pp. 195–201/171–76). Because of the conditions under which the principles are chosen and implemented, "just institutions are collectively rational and to everyone's advantage" (*TJ*, p. 567/497), and are recognized as such. But that does not mean that the person who considers cheating on his taxes is irrational, or that he is overcome by a blind passion for money. Rather, when he knows that others will pay their taxes, he reasons—as Rawls says—that "even though the marginal social value of his tax dollar is much greater than that of the marginal dollar spent on himself, only a small fraction thereof redounds to his advantage" (*TJ*, pp. 336/295–96). "From a self-interested point of view" (*TJ*, p. 336/295), he sees that when others pay their taxes, he is better off cheating on his taxes than paying them. And so the balance of his reasons in this case seems to tilt against his sense of justice and in favor of shirking, even when others do their part.

Thus from the "self-interested point of view," it is rational for the individual to defect from a collectively rational arrangement. Stability is threatened because what is rational for one person who adopts the "self-interested point of view" is rational for everyone else as well: if each person thinks others will act justly, then every person's balance of reasons seems to tilt against acting justly himself. Of course, if no one else is paying taxes, then there will not be enough money to fund social programs even if I pay mine. In that case, at least from a "self-interested point of view," I cannot expect any benefit to redound to me from paying taxes. The payoff of not paying taxes when no one else is paying theirs exceeds the payoff of paying them, and so my best response to others' unjust behavior is to refuse to do my duty. Shirking therefore seems to be my best course of action, regardless of whether others comply with or defect from the terms of cooperation. But what is true of me is true of everyone else. So shirking seems to be everyone's best strategy, regardless of what others do. The stability of justice as fairness is therefore threatened, not by irrationality, but by conduct that seems, on balance, to be in the rational interest of each individual when he adopts the "self-interested point of view."

The payoffs for the strategies of paying and not paying are depicted in Table II.1. I have depicted this as a two-person game for simplicity's sake, with payoffs for player 1 given first in each box and where '~Full benefits' means 'approximately Full benefits'.

Table II.1

		Player 2	
		Act justly	Act unjustly
Player 1	Act justly	Full benefits - taxes paid Full benefits - taxes paid	~Full benefits - taxes paid ~Full benefits + taxes saved
	Act unjustly	~Full benefits + taxes saved ~Full benefits - taxes paid	No benefits + taxes saved No benefits + taxes saved

Despite the inaccuracies that result from depiciting a multiparty public goods game as a two-party game (reflected in the vagueness of "approximately Full benefits"), I take it that this analysis of public goods problems is familiar enough. The table shows a prisoner's dilemma. Thus, the second threat to the stability of justice as fairness is that it will be undermined by a "generalized prisoner's dilemma" (cf. *TJ*, p. 577/505). The fact that in the WOS, everyone's balance of reasons tilts against being just when he adopts the "self-interested point of view" does not itself imply that everyone in the WOS—or that very many people in it—will be unjust. Each member of the WOS has a sense of justice, and none may accede to the temptations of the "self-interested point of view." But clearly, the stability of justice as fairness would be more secure if the temptations present in that point of view were removed.

One way to remove the temptations is to alter the payoffs for the two courses of action open to the players, and with it each individual's balance of reasons, so that acting from Rawls's principles is *always* in each person's best interest—or, more formally, so that acting from those principles is each person's dominant strategy. But the alteration in payoffs would have to be very great if acting from those principles is the best response to injustice. Replying to injustice in this way would leave someone open to exploitation and hence to very great loss, and the compensation for that loss—or the loss for replying unjustly oneself—would have to be very great to tip the balance in favor of Rawls's principles regardless of the behavior of others. Moreover, as we shall see in §VI.2, such a dramatic alteration of the payoffs is not needed to remove the second threat to stability, since that threat arises on the assumption that others will act justly. All that is necessary is to alter the payoffs so that acting from Rawls's principles is each person's "best reply" (*TJ*, p. 568/497) when others behave similarly.

I have depicted the altered payoff in Table II.2, in which payoffs are changed by some perceived loss for defection.

Table II.2

		Player 2
	Act justly	Act unjustly
Player 1		
Act justly	Full benefits − taxes paid Full benefits − taxes paid	~Full benefits − taxes paid (~Full benefits + taxes saved) − loss incurred
Act unjustly	(~Full benefits + taxes saved) − loss incurred ~Full benefits − taxes paid	(No benefits + taxes saved) − loss incurred (No benefits + taxes saved) − loss incurred

If the loss is great enough, then each of the players disvalues it more than he values the money he can gain by cheating on his taxes. Then even from a "self-interested point of view," each player sees that he is better off acting justly when others do. Each player's balance of reasons tilts in favor of being just when others are also just, and so acting justly is each player's best response to the just conduct of others. The state of affairs in which everyone responds justly to the justice of others is therefore a Nash equilibrium and the second threat to stability is removed.[9]

What of the first threat to stability? That threat arises on the assumption that each person wants to act justly, but needs the assurance he will not be taken advantage of. Since a WOS is a just society, everyone is already behaving justly, so what each person needs to be assured of is that others will continue to act justly rather than defect. Suppose that each person knows everyone else's balance of reasons tilts in favor of acting justly when others do, even from the "self-interested point of view." Then each knows that no one else has sufficient reason to take advantage of him and the *mutual assurance problem* is solved. No one has sufficient reason for preemptive defection. Thus—assuming the special circumstances of a WOS—the way to avert the first threat to stability is to alter each person's payoffs so that everyone knows the second threat is averted (see *TJ*, p. 336/296). The society will then be in a just equilibrium. And if each person knows that everyone else's balance of reasons tilts in favor acting from her sense of justice *whenever* she adopts the "self-interested point of view," then it will remain in equilibrium. Then the "hazards of the generalized prisoner's dilemma are removed by the match between the right and the good" (*TJ*, p. 577/505). "No tendencies to instability exist" (*TJ*, p. 567/497) and justice as fairness is "as stable as one can hope for" (*TJ*, p. 399/350).

The crucial step in bringing about stability is therefore bringing about the requisite "match between the right and the good." But how can that be done in

9. A strategy-combination is a Nash equilibrium if each player's strategy is his best reply to the strategy played by the others.

each person's case? And how can the fact that the match exists in each case become an object of public knowledge?

Hobbes's work illustrates one way of achieving stability by bringing about the match. It will prove useful to consider Hobbes here because his work brings out the contrast I want to draw between inherent and imposed stability. Hobbes thought that the state of nature would have the structure of a generalized prisoner's dilemma, as Rawls himself notes (*TJ*, p. 269/238). It might be possible to identify fair terms of cooperation that would make everyone better off than he would be in that condition. Hobbes famously thought that people would not honor those terms on their own. But if they sufficiently disvalue punishment and fear, then a sovereign's addition of punishment and fear to each person's payoff table will transform his table from the first to the second.[10] Then when each person adopts the "self-interested point of view," he will see that when others cooperate, he is worse off shirking his duty than doing it.[11] If each player also knows that everyone else attaches sufficient disvalue to punishment and fear, and if each knows that the sovereign deploys these measures effectively, then the changes in each person's payoff tables are a matter of mutual knowledge. A sovereign with the absolute power to punish, and who is known to be effective over time, can therefore stabilize the terms of cooperation in a Hobbesian society.

Sanctions are needed because the terms of cooperation in a Hobbesian state would not generate their own support. More precisely, the terms of cooperation among citizens, and the institutions designed to implement and publicize those terms, do not themselves alter the payoff tables and remove the second threat to stability. That threat must be removed "by forces [*outside*] the system" of social cooperation—by a sovereign who is not himself a subject and does not himself have a payoff table, but who stands above the subjects and who alters their payoff tables using coercion and threat. That is why, when contrasting his own treatment of stability with Hobbes's, Rawls observes that "One way of interpreting the Hobbesian sovereign is as an agency *added to* an unstable system of cooperation in such a way that it is no longer to anyone's advantage not to do his part given that others will do theirs."[12]

10. For a clear statement of this way of avoiding prisoner's dilemmas, see Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984), pp. 133–34. For interpretations of Hobbes that impute this function to the sovereign, see Jean Hampton, *Hobbes and the Social Contract Tradition* (Cambridge: Cambridge University Press, 1986), pp. 132ff. and Edna Ullmann-Margalit, *The Emergence of Norms* (Oxford: Oxford University Press, 1977), p. 67.

11. Thus in the "Introduction" to *Leviathan*, Hobbes says "reward and punishment" are the nerves of the Leviathan "by which fastned to the seate of the Soveraignty, every joynt and member is moved to perform his duty." See Thomas Hobbes, *Leviathan* (New York: W.W. Norton, [1651] 1997), ed. Richard E. Flatham and David Johnston.

12. Rawls, "Sense of Justice," *Collected Papers*, p. 104 (emphasis added). Interestingly, in the lecture on Hobbes devoted to "The Role and Powers of the Sovereign," Rawls writes as if Hobbes thought the sovereign was needed only to solve what I have called the *mutual assurance problem*; see John Rawls, *Lectures on the History of Political Philosophy* (Cambridge, MA:

The Hobbesian sovereign and the penal system are presumably part of the basic structure of a Hobbesian society. If we liked, we could still say that the *basic structure* of a Hobbesian society stabilizes itself, and that that *society* or its *governing apparatus* is inherently stable. But doing so would blur the contrast that Rawls tries to draw between his own account of stability and Hobbes's, for it would obscure the fact that in a Hobbesian society, the *terms of cooperation* are not inherently stable. A Hobbesian society might be stably just, but if so, the stability of its conception of justice would be *imposed* rather than *inherent*.

§II.2: Matching the Right and the Good in Justice as Fairness

To see how the "match between the right and the good" is brought about in justice as fairness, note that so far, I have discussed the second threat to stability as if the choices open to members of the WOS were *actions*. I have done so for ease of exposition. I have also done so because the choice of how to act is the choice we most frequently face, because the possibility of free-riding is so pervasive and familiar a threat to the stability of fair cooperative schemes of all kinds and because it is a threat with which Rawls is quite naturally concerned. But this exposition can suggest that Rawls tries to avert the threat by assuming that members of the WOS treat each act independently, and by showing that each time they have to act, they weigh their reasons and determine that the balance tilts against free-riding. It can suggest, that is, that he tries to show that in every encounter with others, just action is the "best reply" (*TJ* p. 568/497) to the just action of others, while treating the encounters as independent.

But this is an implausible way to avert the threat of instability for two reasons. One is that trying to avert the threat this way would require Rawls to confront all the complications of iterated games among a large number of players, many of whom may not recognize each other from previous encounters because of the frequency and anonymity of interactions among citizens in a large society. These are problems that are said to beset Hobbes's own solution to the stability problem, and to beset contemporary Hobbesian accounts.[13] Moreover, if we imagine people considering each act in isolation from the others, we ignore the effect of past choices on character, and ignore the effect of character on what

---

Harvard University Press, 2007), ed. Samuel Freeman, pp. 78–79. In Appendix A to that lecture, however, he seems to recognize that Hobbes thought the sovereign also has the role of changing subjects' payoff tables; see p. 91, paragraph 4(b).

13. For incisive discussions of some of the difficulties besetting David Gauthier's Hobbesian account, see Gregory S. Kavka's review of Gauthier's *Morals by Agreement* (New York: Oxford University Press, 1986) in *Mind* 96 (1987): pp. 117–21; also Alan Nelson, "Economic Rationality and Morality," *Philosophy and Public Affairs* 17 (1988): pp. 149–66. More generally, see Samuel Freeman, "Reason and Agreement in Social Contract Views," *Philosophy and Public Affairs* 19 (1990): pp. 122–57.

payoffs various courses of action will seem to promise. Someone who has a history of behaving justly may value certain returns of a just action that the unjust person entirely discounts. Where the just person may count virtue as its own reward, the unjust person may think virtue a reward that is not worth having. This opens the possibility of a different way averting the threat of a prisoner's dilemma, one that is more psychologically plausible and that enormously simplifies the problem of matching the right and the good.

To illustrate the solution, consider another problem discussed in the literature of game theory, the mortarmen's dilemma.[14] In its simplest form, the problem concerns two machine gunners, both of whom must remain at their posts to stop an enemy advance and save their city. Because they are an advance line of defense, their positions are dangerous. If both remain in their posts, then they have a 50–50 chance of being captured by the enemy. If one defects while the other remains, the defector will escape with his life and go home while the one who remains faces a 90% chance of being killed. If both desert, and the enemy overruns their positions without resistance, then they face a 70–30 chance of being captured. The collectively rational solution is for the two to remain at their posts, but the rational choice for each individual is to desert. The mortarmen's dilemma is therefore a prisoner's dilemma.

Suppose that the mortarmen's unit tries to counter desertion and cowardice by fostering a sense of honor, so that each member of the unit is motivated to do his duty. Even so, their commander may worry that the mortarmen will be tempted to act against their sense of honor and to desert in the heat of battle. We might counsel the commander to change the payoffs, perhaps by attaching very strong sanctions to desertion, so that desertion becomes the less attractive choice in each instance. Or we might argue that the danger will not arise because the mortarmen see a life governed by a code of honor as part of their good, at least when they know that others are governed by it as well, so that each sees a life of honor as the best response when they know others commit to such a life as well.

The latter strategy is like the one Rawls employs to avert the hazards of the generalized prisoner's dilemma. To show that, I need to say more than I have so far about the sense of justice.

As I said in the previous section, Rawls argues that the institutions of the WOS foster a sense of justice in those who live under them. The sense of justice, like the sense of honor, is an established disposition to judge and act from principles of right conduct, and to make amends for violating them, provided others are similarly disposed. Suppose that someone in the WOS—call her Joan—has a sense of justice and regularly feels its promptings. Those prompting include many desires, including the desire to do the right thing, the desire not to do wrong and the desire to seek forgiveness when she has done wrong. We can sum this up by saying that Joan's sense of justice is a desire to be just. Suppose, furthermore, Joan's desire to be just is not only a desire about her actions. It is also a highest-order desire, a desire about all her other desires. We

14. See Ullmann-Margalit, *Emergence*, pp. 31ff.

may suppose that because she wants to be just, she is troubled by the desires she sometimes feels to act unjustly and wishes she did not have them. Rather, she is, or tries to be, the kind of person who attaches little value to what she can gain by acting against the principles of justice. If she sees that an action would be contrary to the principles, she is, or tries to be, the kind of person who does not even consider performing it, at least when she knows others are just. If she succeeds in being and remaining that kind of person, then Joan's life is ruled by her highest-order desire to act from the principles or, as I shall say, whose desire to act from the principles is a highest-order desire that is *regulative of her life*. Thus, if Rawls can show that each member of the WOS would indeed be like Joan, and would be known to be like her, then he could show the stability of the WOS while avoiding the moral and philosophical difficulties with imposed stability and repeated games.

Showing that everyone in the WOS would normally acquire a sense of justice is not enough to show that everyone would be like Joan. For even if a sense of justice is successfully cultivated, that disposition can itself still be undermined by temptations that arise from within the "self-interested point of view" (*TJ*, p. 336/295). Members of the WOS may resent their own sense of justice because of its costs. Once they realize that their society is set up to encourage that sentiment, they may worry that they have been illegitimately indoctrinated. Even if they do not try to extirpate their sense of justice, they may wonder what place it is rational for them to give that disposition in their plans of life. Wouldn't they regret allowing it to regulate their lives, so that they act justly *on principle*? Shouldn't they treat their sense of justice as one desire among others, deciding whether to act justly case-by-case?

From "the self-interested point of view," the latter may seem the more rational course of action, since the former is a commitment to forego the gains of injustice, while the latter leaves one free to choose the action—including free-riding—which promises the greatest expected gains. These worries and temptations cannot simply be assumed away, any more than we can assume that the mortarmen will not face them. Like the mortarmen, the members of the WOS need to be convinced that their settled disposition to do the right thing is part of their good, and they need to be convinced that this is so even from the "self-interested point of view." While they need not be convinced that it is good to act from the principles of justice regardless of what others do, they need to be convinced that it is good to be just when others are just as well.

This means that though Rawls can avoid the problems with repeated games by having the players choose between two lives they might lead, he is left with another game-theoretic problem, this time one in which the strategies open to players are policies rather than actions. That is why Rawls implies that the problem he still faces, even after showing that everyone in the WOS would have a sense of justice, is *not*—as I have so far supposed—that of showing that just action is each person's "best reply" to the *just actions* of others. He still has to show that, even from the "self-interested point of view," "the *plan of life* which [is regulated by the sense of justice] is [each person's] best reply *to the similar plans*

of his associates" (*TJ*, p. 568/497, emphasis added). And he still has to show that each can be sure others in the WOS do in fact have similar plans. Thus both of the threats to stability that I identified in the last section—temptations that arise from the "self-interested point of view," and the *mutual assurance problem*—arise with respect to each person's policy of, or his commitment to, being just.

The mortarmen's dilemma helps to state the challenge more formally. The mortarmen are governed by a code of honor. If the code is to be inherently stable, then the practices that implement the code must encourage a sense of honor among soldiers. They must also bring it about that each does what he must to maintain his sense of honor. They do that by bringing it about that each sees—and knows that all the others see—a life of honor as a better life than a life in which each he tries to root out his sense of honor or becomes the kind of person who judges case-by-case whether to stand fast or desert his comrades. And they must bring it about that each soldier sees this even from a "self-interested point of view." Thus those practices must bring it about that each soldier sees himself—and knows that every other sees himself—as faced with a payoff table that resembles Table II.2 rather than Table II.1, but one in which the cooperative strategy is not an action, but a commitment to maintaining his sense of honor over the course of his service.

Similarly, showing the inherent stability of a conception of justice requires showing that the institutions which implement it stabilize themselves in two ways:

> First, they must elicit a sense of justice.
>
> Second, they must themselves bring it about that even when each member of a just society assesses his reasons from a "self-interested point of view," he still sees that the balance of his reasons tilts toward maintaining a supremely regulative sense of justice—rather than deciding whether to be just case-by-case—when others do so as well.

Then, once the *mutual assurance problem* is solved, it will be rational for each member of the WOS to preserve his own sense of justice, and the WOS will be stably just.

Rawls shows how institutions stabilize themselves in the first way in chapter 8 of *TJ*. To show that they stabilize themselves in the second way, he needs to show something about the payoff table that each member of the WOS takes himself and others to be faced with when in the "self-interested point of view." He needs to show that institutions bring it about that in the "self-interested point of view," each takes himself to be faced with, and knows that everyone else takes himself to be faced with, not the payoff table of a prisoner's dilemma, but payoffs like those shown in Table II.3, where $A > B > D > C$. In that case, each person sees—from the "self-interested point of view"—that it is better for him to be a just person when others are also just, and knows that everyone else sees that it is better for her to be just as well. Though I shall add some important qualifications, it is useful to think of Rawls as trying to do roughly that.

Table II.3

	Player 2	
	Maintain regulative desire to act from the principles	Decide case-by-case
Maintain regulative desire to act from the principles	A, A	C, B
Player 1		
Decide case-by-case	B, C	D, D

Even after Rawls establishes that each member of the WOS would face a payoff table like Table II.3, he still needs to solve the *mutual assurance problem*. He recognizes that in a large society like the WOS of justice as fairness, members cannot all be acquainted with one another. We shall see that providing each person the assurance that each of the others will preserve and act on his sense of justice may require the existence of coercive institutions, just so that each knows those who are not otherwise inclined to act justly will be deterred from acting unjustly. But Rawls intimates that these institutions have a very different function than penal institutions in a Hobbesian society. In a Hobbesian society, they exist to shift each person's balance of reasons. In the WOS of justice as fairness, they exist only to clinch the solution to the *mutual assurance problem* (*TJ*, p. 269/237). Unfortunately, Rawls does not say enough about this, and I shall return to the point in §VII.10.

The technicalities of stability mask the extraordinary ambition of Rawls's attempt to demonstrate the inherent stability of justice as fairness. It is the task of showing that principles of justice which are collectively rational are also, when institutionalized, "self-reinforcing" and so are immune to the instability that results from collective action problems. They reinforce themselves by bringing it about that each sees adhering to them voluntarily over the course of life to be part of her good.[15]

Let me bring out the magnitude of this task by redescribing it in terms suggested by Samuel Freeman's treatment of stability in Rawls.[16] The collective rationality of the principles is shown by their adoption in the OP. Given the special conditions of the OP, the fact that the principles would be adopted there shows that they are principles we would give ourselves. To show that members of the WOS would acquire a sense of justice shows that they would comply with those principles voluntarily or freely. It would show, we might say, that members of the WOS would act *from*, and not merely in accordance with, principles they would give themselves. Showing the inherent stability of justice as fairness requires showing that each would voluntarily do and be

15. Rawls, "Sense of Justice," *Collected Papers*, p. 106.
16. See Freeman, "Congruence and the Good of Justice."

known to do what was necessary to preserve his sense of justice, so that the justice of a WOS over time could be maintained by the autonomous activity of its members.

But even if the right or the most philosophically defensible conception of justice can be identified, and be shown to be collectively rational, why should we think that enough people will accept and act from it that a society regulated by it will remain just? Is the disposition to act from it a disposition to act morally, according to the most philosophically defensible account of moral motivation? And even if it is, why should we think people can develop and sustain that motivation, so that they act justly of their own volition? Showing that they would seem to require showing that being a just person fits with the deepest and most powerful motivations of our human nature. Why should we think that it is, especially when human history seems to provide such powerful evidence to the contrary?

As we have already seen, Hobbes thought it was not, and argued that the second condition of inherent stability cannot be met. I believe many other thinkers in the history of philosophy would agree. I have used Hobbes to illustrate the distinction between inherent and imposed stability because Rawls himself seems to use Hobbes that way and because the coercion exercised by the Hobbesian sovereign may seem the clearest instance of a stabilizing "force [outside] the system" of cooperation. But inherent stability can also be distinguished from stability—of a state or of a conception of justice—that is achieved by the widespread acceptance of false beliefs, such as Plato's Noble Lie, or that is achieved by the widespread acceptance of a single religion. For acceptance of these systems of belief may also encourage a sense of justice and change the way citizens think their balances of reasons tilt. If they all accept a single religion that makes salvation conditional on obedience to the powers that be, for example, they may all think that even from a "self-interested point of view," the balances of their reasons tilt in favor of obedience. Indeed, as Rawls notes, philosophical or religious uniformity was long thought to be necessary for stability.[17]

Stability achieved through such uniformity may not seem to be imposed, for the distinction between inherent and imposed stability depends upon a distinction between stabilizing forces that are inside and outside "the system of cooperation." Since ideologies might not seem to be outside the system, stability that depends upon them might seem to fall on the wrong side of the inherent-imposed distinction. But the thought that they do depends upon the assumption that the Noble Lie or the single religion would be accepted voluntarily. Rawls is surely right when he remarks in *PL* that "in the society of the Middle Ages . . . the Inquisition was not an accident; its suppression of

17. John Rawls, "The Priority of Right and Ideals of the Good," *Collected Papers*, pp. 449–73, p. 464.

heresy was needed to preserve…shared religious belief" (*PL*, p. 37). The universal acceptance of either the Noble Lie or a single religion would require the oppressive use of state power. Like Hobbes's sovereign, Dostoevsky's Grand Inquisitor may therefore be interpreted "as an agency added to an unstable system of cooperation" to bring about stability.[18] If Hobbesian stability is imposed, then so is stability that depends upon the Noble Lie or upon adherence to a single religion.

Thus Hobbes, Plato, and many other political philosophers have been concerned with *some* questions of stability. Perhaps one of the questions that concerned them was that of how society could be stably just. But because Hobbes argued that stability had to be imposed and because Plato and most other philosophers have resorted to stabilizing mechanisms which would have to be, the problem that concerned them was very different from that of showing that a conception of justice is inherently stable. At the beginning of §II.1, I quoted Rawls's seemingly curious remark that the problem of stability "has played very little role in the history of moral philosophy" (*PL*, p. xix). Now that we see Rawls's concern with *inherent* stability, we can see what he meant, for the problem of showing inherent stability is one that many philosophers have thought insoluble. On my reading, Rawls thought it was not. He wanted to show that justice as fairness is inherently stable, and he tried to do so by showing that the institutions of the WOS would stabilize themselves in the two ways listed above.[19]

§II.3: Congruence and Stability

This reading of *TJ*, according to which Rawls's treatment of stability takes up game-theoretic concerns, may strike the reader as rather novel, since this is hardly the usual way of reading Rawls. I have tried to support the interpretation by drawing together hints Rawls drops at various places in his published work—such as his distinction between two sources of instability that exist

18. The quoted passage is from Rawls, "Sense of Justice," *Collected Papers*, p. 104. For the Grand Inquisitor and the Noble Lie, see *TJ*, p. 454, note 1/398, note 1.

19. The second part of *TJ*, on institutions, has received very little commentary despite the fact that it comprises approximately a third of the book. I believe readers often assume that Rawls devoted so much attention to the subject because he thought it important to show how justice as fairness could be implemented. This is a natural enough assumption to make, given one of Rawls's remarks about the purpose of part II (*TJ*, p. 95/81). But if my reading is correct, that part has another purpose as well. Rawls wants to show something vitally important about justice as fairness—namely, that it is stable. He says that he can show that it is stable by showing something about the institutions that satisfy it—namely, that they "generate their own support." While he shows *that* in chapters 8 and 9 of *TJ*, the chapters on institutions are needed to supply premises for the arguments of those later chapters.

even after members of the WOS have been shown to have a sense of justice (*TJ*, pp. 336/295–96), his passing mention of "the hazards of the generalized prisoners' dilemma" (*TJ*, p. 577/505), his intimation that justice as fairness would be "inherently" (*TJ*, pp. 144/125, 498/436) or "intrinsically"[20] stable, and the grounds on which he contrasts his own treatment of stability and Hobbes's (*TJ*, p. 497/435).

Rawls hints at his game-theoretic concerns most obviously when he says he wants to show that a commitment to acting justly is each citizen's "best reply to the similar plans of his associates" (*TJ*, p. 568/497). This is a clear indication that Rawls wants to show that a state of affairs in which everyone regulates his plans by terms of cooperation is a Nash equilibrium. The interpretation derives some additional support from the fact that crucial elements of it have been seen by others. For example, Edward McClennan explains the stability problem in *TJ* in an especially clear and illuminating way because he sees the key distinction between imposed and inherent stability.[21] But the best way to substantiate this reading is to show how Rawls's arguments for stability in *TJ* actually respond to these concerns.

I have already remarked that in *TJ*, Rawls sets up the problem of stability as a two-stage problem (*TJ*, p. 453/397). In the first stage, he shows the first thing that I said needs to be shown if justice as fairness is to be shown inherently stable: that members of the WOS would all acquire, and know that others would acquire, a sense of justice. The second part of the stability problem is that of showing that the right and the good are *congruent*.

Rawls does not define congruence in *TJ*, and his remarks about it are difficult to interpret. Congruence is clearly a relation, but Rawls does not say clearly just what the relata are. It would be natural to think that congruence is a relation that holds, in the first instance, between the right—understood as the principles of justice or justice as fairness—and *each person's good*, so that congruence obtains, as it were, person-by-person. I do not think that this interpretation is correct; sustaining it would, I think, force subtle misreadings of important passages. Instead of starting with Rawls's texts, I want to present my own interpretation of congruence by returning to Joan, the member of the WOS whom I introduced in the previous section. I shall then try to show how this interpretation squares with Rawls's text.

Like all of us, Joan makes plans for her life. In making those plans, Joan reflects on and tries rationally to schedule the satisfaction of her longer- and shorter-term aims and desires. Clearly, Joan may be tempted by plans or subplans that conflict with the demands of justice. She may, for example, be tempted to cheat on her taxes because she wants extra money to spend or to pass along to her children or to give to her favorite charity. Since Joan is a

20. Rawls, "Sense of Justice," *Collected Papers*, p. 106.

21. Edward McClennan, "Justice and the Problem of Stability," *Philosophy and Public Affairs* 18 (1989): pp. 3–30, pp. 7–8.

member of a WOS, she has a sense of justice. And so when she surveys her desires and aspirations, makes her plans and asks how she wishes to live, she has to decide how highly she values her desire to be just, whether to maintain it and what place that desire has in their plans.

Rawls says in *TJ* that "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*, p. 409/359). What does it mean to say that a plan of life "establishes" a "point of view"? I believe what Rawls has in mind is this: When Joan makes various kinds of judgments and decisions, she does so on the basis of certain desires, bodies of information, and canons of reasoning. Points of view are given by the desires, rules of reasoning, and information someone draws on when she makes decisions or renders judgments of the relevant kind. If this is right, then when Joan make her plans, she makes them from within some point of view. In that point of view, she draws on all the information then available to her about what she wants, what resources she has available, what the future may be like, how others will respond to her action, and where she is in the ongoing execution of plans she has already made. It is because Joan makes judgments of value from within plans already made that Rawls says the point of view from which those judgments are made is established by her plan of life. Since Joan reasons using the rules of what Rawls calls "full deliberative rationality" (*TJ*, p. 408/359) in that point of view, I shall refer to the point of view from which Joan draws up her plans as the "viewpoint of full deliberative rationality."

The questions of whether to maintain her sense of justice and what place to give it are questions Joan answers from this point of view. Joan's sense of justice is a desire to act from the principles of justice for their own sake, and to give them priority in her practical reasoning. So as Joan makes her plans from within one viewpoint, using principles of rational choice, she has to ask herself what place or weight she gives to her desire to act from another set of principles. If, when Joan adopts the viewpoint of full deliberative rationality, it is rational for her to maintain her sense of justice as a highest-order regulative desire, then there is a "match" between the principles of full deliberative rationality and the principles of justice. Planning with one set, she affirms the other.

Of course, whether it is rational for Joan to maintain her sense of justice as a supremely regulative desire depends upon what else she wants. The "match" between the two sets of principles is conditional on the content of Joan's desires. But now suppose that Joan is a typical member of the WOS in this sense: the desires that move her to treat her sense of justice as supremely regulative are desires that everyone in the WOS has. If we assume that Joan is also typical in the weights she attaches to those desires, then it will be rational for everyone in the WOS to decide to treat his sense of justice as regulative when he adopts the viewpoint of full deliberative rationality. In that case, the match between sets of principles is not conditional on the idiosyncratic

content of any given person's desires. There is simply a match "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*, p. 514/451). This match between sets of principles is congruence (see *TJ*, p. 514/451).

Rawls also says that congruence is a "match between justice and goodness" (*TJ*, p. 399/350). If congruence is taken to be a match between justice and any given person's good, then the word "goodness" will seem out of place in this remark. But if we interpret congruence as I have, this remark is a perfectly understandable piece of shorthand. It expresses, albeit pithily, the claim that congruence holds between what would be chosen in the OP and the desires that are part of a rational plan of life, as adopted in the viewpoint of full deliberative rationality. What is chosen in the first point of view is justice; according to Rawls's theory of goodness as rationality, what is chosen in the second point of view is goodness.

The viewpoint of full deliberative rationality and the original position are points of view we can adopt when reasoning practically. In his later work, Rawls distinguished two moral powers, which he called the *Rational* and the *Reasonable*. The *Rational* is our capacity for a conception of the good; the *Reasonable* is our capacity for a sense of justice. The two points of view – the viewpoint of full deliberative rationality and the OP—are associated with these two moral powers. They are, we might say, points of view within practical reason.

The prisoner's dilemma is sometimes described as a "paradox of rationality" because it shows that individual and collective rationality—understood as the rational pursuit of individual and collective interests—can conflict. The existence and intractability of the paradox has led some thinkers to question whether the conception of rationality at work in setting up the prisoner's dilemma is the right one.[22] Rawls thinks an agreement reached in the OP is collectively rational (*TJ*, p. 567/497). If that agreement were undermined by a collective action problem, it would be a particularly disturbing paradox of this kind, one that would raise similarly pressing questions about Rawls's conception of practical reason. Since the OP is associated with one power of practical reason and the viewpoint of full deliberative rationality is associated with another, the vulnerability of an agreement reached in the OP to the "hazards of the generalized prisoner's dilemma" (*TJ*, p. 577/505) would show that the constituents or elements of practical reason can be at odds. Moreover, it would show that they can be at odds over the justice of the basic structure, where a collectively rational solution is urgently needed. This possibility would raise doubts about whether the distinction between the Reasonable and the Rational accurately maps the psychological terrain.

22. Amartya Sen, "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs* 6 (1977): pp. 317–344.

If congruence obtains and the sense of justice is treated as supremely regulative over time, then one point of view within practical reason—the viewpoint of deliberative rationality—is subordinated to the other. As we shall see, Rawls thinks that this subordination of one point of view to another unifies practical reason. To show congruence in the WOS is therefore to show that, in the conditions of the WOS, practical reason itself has a kind of unity. Showing this removes the doubts that a paradox of rationality would raise. Just what kind—or kinds—of unity practical reason has is a question I shall defer until Chapter VII. For now, suffice it to say that that unity is realized or exhibited in the ongoing life of the just person. To live as a just person is, Rawls thinks, to live a life in which the powers of practical reason are unified.[23] We shall see in Chapter VII that Rawls thinks exercising one's faculties of practical reason by acting as a just person, unifying those faculties by taking the sense of justice as supremely regulative, is part of what makes being a just person "a leading human good" (*TJ*, p. 426 note 20/374, note 20).

I have said that congruence is a relation that holds between sets of principles, rather than between the principles of justice and anyone's good. But if congruence does not consist in a relation between justice and anyone's good, it still has *implications* for each person's good. Rawls says congruence "implies that members of the well-ordered society"—by which I take it he means "*each and every* member" or "*all* members"—"when they appraise their plan of life using the principles of rational choice, will decide to maintain their sense of justice as regulative of their conduct toward one another" (*TJ*, pp. 514/450–51). If each person makes a rational decision to maintain his sense of justice as regulative of his plan, then it must be because when each member of the WOS assesses her reasons from the viewpoint of deliberative rationality, she sees that the balance of all her reasons—self-interested and not—taken together tilts toward maintaining it.

Thus on my reading, the problem of showing that congruence obtains is that of showing that each member of the WOS sees—when she adopts the viewpoint of full deliberative rationality—that her balance of reasons tips toward maintaining a supremely regulative desire to be just, and draws up her plans accordingly. More precisely, it is, in the first instance, the problem of establishing what I shall call the *Congruence Conclusion* or C_C:

23. William Galston once criticized Rawls for relying on two "not wholly consistent" accounts of motivation: moral motivation and "the narrowly self-interested rational calculat[ion] of modern economic and social choice theory"; see William Galston, "Moral Personality and Liberal Theory: John Rawls's 'Dewey Lectures,'" *Political Theory* 10 (1982): pp. 492–519, p. 493. Galston is surely correct to note that there is some tension between the two kinds of motivation he distinguishes, but that is because there is a similar tension within practical reason. Just how that tension is to be resolved, so that we affirm our sense of justice and unify our practical reason, is precisely what the treatment of congruence is supposed to show.

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

Clearly if Rawls can show that this conclusion holds whenever members of the WOS consider whether to preserve their sense of justice, then—having shown that members of the WOS would normally acquire a sense of justice—he can show that the WOS would be stably just.

When I laid down the conditions of inherent stability, I said that members of the WOS must decide to maintain their sense of justice even when they adopt what I called—following Rawls—a "self-interested point of view" (*TJ*, p. 336/295). Simply establishing the *Congruence Conclusion* does not establish that, since the *Congruence Conclusion* refers to the viewpoint of full deliberative rationality and someone who adopts that viewpoint takes account of *all* her ends, including the ends associated with her sense of right. A proof of congruence will be especially powerful, and will show what is needed for inherent stability, if it gets to the *Congruence Conclusion* by showing that in the WOS, each person's balance of reasons would tip toward maintaining his sense of justice even if he were not moved by the desire to be just for its own sake.

The "self-interested point of view" is not the point of view from which human beings typically reason. We can, however, fall into it or adopt it on reflection. In the course of ordinary life, we may find ourselves asking what decisions we would make if we were ultimately moved only by our self-interest or did only what we want to do, while construing "want" narrowly. If members of the WOS would find it rational to maintain their sense of justice even then, then they will plan to maintain their sense of justice when they adopt the viewpoint of full deliberative rationality. For the only difference between that viewpoint and the "self-interested point of view" is that in the latter, we value the ends of justice only to the extent that securing them gets us other things we want. Furthermore, the fact that members of the WOS would find it rational to maintain their sense of justice from a "self-interested point of view" reflects a fact about the coherence—the congruence—of their reasons. It reflects the fact that the reasons telling against being a just person are not strong enough to undermine the sense of justice. Rather, even the desires of self-interest "pull" members of the WOS toward justice and are satisfied when they live maintain their sense of justice. This gets Rawls to the *Congruence Conclusion* and to the stability of the WOS.

We shall see later that Rawls introduces these arguments with considerably more refinement than I have so far. My use of payoff tables suggests that each person's balance of reasons depends upon the availability of cardinal measures for the benefits of two strategies. As we shall see in §VI.3, one of the most ingenious elements of Rawls's argument for congruence is the way he establishes a conclusion about each person's balance of reasons without

supposing—implausibly—that cardinal measures are available. Moreover, what I have referred to as the "self-interested point of view" (*TJ*, p. 336/295) is itself refined, and becomes the point of view of "a person following the thin theory of the good" (*TJ*, p. 569–70/499). It is very important that this point of view is *not* that of a self-interested person, on the usual understanding of "self-interested"—a point I shall make clear when I discuss the thin theory in Chapter III. Finally, Rawls assumes that the person who has a sense of justice has a very different set of values and ends than the person who does not. This difference is not just a matter of the just person's valuing the ends of justice for their own sake. Rather, a sense of justice, Rawls thinks, has far-reaching effects on the character of the just person and affects what he takes his balance of reasons to be, even when he judges according to the thin theory.

I shall begin looking at Rawls's arguments for congruence in Chapter VI. Because I look at the arguments in that chapter, I shall refer to one of the claims for which Rawls argues as 'C_6.' As Rawls states the claim, it says "it is rational for someone, as defined by the thin theory, to affirm his sense of justice" (*TJ*, p. 568/497). I shall put the claim somewhat more precisely, to fit with argument Rawls offers for it and to show that it is a stronger variant of C_C, the *Congruence Conclusion*. As I shall word it, the claim is:

> C_6: Each member of the WOS judges, *from within the thin theory of the good*, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

But C_6 seems to be a very strong claim. It seems to imply that each person will judge that it is in his interest to act justly as a matter of principle, quite apart from his desire to be just, and that it is in his interest do so *regardless of how he is treated by institutions and by other people*. Someone who made justice a policy, come what may, might be very admirable but he would also very vulnerable to losses of all kinds. We might wonder whether anyone could judge that it is rational to leave himself so vulnerable, particularly if he renders the judgment while supposing that he is not ultimately moved by the goods of justice.

It is important that Rawls's argument for the *Congruence Conclusion* does not depend upon so strong a claim. He remarks in one place that "even with a sense of justice men's compliance with a cooperative venture is predicated on the belief that others will do their part" (*TJ*, p. 336/296). If someone who is moved by a sense of justice needs to believe that others will do their part in order to do his, then the same is presumably true of him when he follows the thin theory of the good. And so one of the crucial moves in the argument for C_6 and C_C turns on a claim which says that each person's cooperation is conditional on his beliefs about others will do. That claim is that from within the thin theory, "the plan of life [in which the sense of justice is affirmed and maintained as supremely regulative] is [each person's] best reply to the similar plans of his associates" (*TJ*, p. 568/497).

I shall call this claim *TJ's Nash Claim*. Precisely stated, it says:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans, *when the plans of others are similarly regulated.*

TJ's Nash Claim can be illustrated using Table II.3, where A > B > D > C and where payoffs are measured not—as when I first introduced the table—in goods valued from the "self-interested point of view," but in goods valued from within the thin theory.

Establishing *TJ's Nash Claim* shows that the state of affairs in which each person maintains his sense of justice, the state of affairs described by C_6, is an equilibrium state. Moreover, it is an equilibrium state the WOS is actually in.

To see this, note first that the WOS is a society in which each person has a sense of justice. This is a deep-seated desire that "can be changed only gradually" (*TJ*, p. 568/498). Even if circumstances were such that someone would be better off becoming the kind of person who decided case-by-case whether to be just, the transformation would take time and she would be open to loss during the transition. So each person would presumably prefer to preserve her sense of justice and, since she faces Table II.3, she will do so if others will. As we shall see, in the WOS, each knows that everyone else faces Table II.3, just as she does. Each therefore knows that everyone else will preserve her sense of justice if she thinks that others will preserve theirs. And each person thinks others will preserve their sense of justice, since each knows that everyone else has a sense of justice and would prefer to preserve it for the same reason she would prefer to preserve hers. This mutual knowledge solves the *mutual assurance problem*, so each will judge from within the thin theory that it is rational to maintain his sense of justice, just as C_6 says.

Once he gets to C_6, Rawls can move to the *Congruence Conclusion*. For if members of the WOS judge that it is rational to affirm their sense of justice even from within the thin theory, then they will surely judge that it is rational to maintain it when they draw up their plans in the viewpoint of full deliberative rationality, taking account of their desire to be just for its own sake. The WOS is in a just equilibrium. And since *TJ's Nash Claim*, C_6, and the *Congruence Conclusion* are true whenever members of the WOS reason from the relevant points of view, the equilibrium is stable.

It is not surprising that C_N should be a pivotal step in Rawls's arguments for congruence. Rawls conjectures that evolution has endowed human beings with a deep tendency to reciprocity. For that tendency to have endured in reflective creatures, it must have been seen to be conducive to our good. The sense of justice builds on this tendency to respond in kind (*TJ*, p. 494–95/433). It is a disposition to conduct ourselves justly when others are just. If members of the WOS are to judge that that disposition is good for them, and is one they want to preserve, they must see that that desire is, on balance good for them. This is just what a successful argument for *TJ's Nash Claim* would show, and Rawls's inability to establish that claim bulked large among the reasons for his turn to political liberalism.

To get an even clearer understanding of congruence, it is helpful to see what successful arguments for C_N, C_6, and C_C would *not* establish.

First, these conclusions need to be distinguished from the claim that the just person performs just acts with ease, that she lacks impediments to just actions in the form of contrary desires or that the performance of just acts is over-determined by the presence of a desire moving the just agent in the same direction. The conclusion that the right and the good are congruent is not a claim about what goes on in the just person act-by-act. It is a conclusion about a higher-order desire to live as a certain kind of person. It *may* be that Joan's awareness of congruence, or of her own conscious affirmation of her sense of justice, facilitates her performance of just acts. But that conclusion would require additional argument that is not to be confused with an argument for C_N, C_6, or C_C.

It is sometimes supposed that Rawls's treatment of congruence is an attempt to uncover a characteristic motive of just action. Someone might think this if he thought that what the congruence of the right and the good showed was that each just action has some good attached to it which functions as the reliable incentive to do what is right. This reading will be less tempting once the previous distinction is drawn. Even so it is worth emphasizing that Rawls thinks the characteristic motives of just action are the desires associated with the sense of justice. The question of congruence presupposes that the characteristic motives of just actions have been identified, and that members of the WOS have those motives and want to treat the principles of justice as supremely regulative. Rawls's concern once he takes up congruence is a concern to show that members of the WOS would find it rational to preserve their desire to act from the principles.

Finally, let me anticipate a point to which I shall return in §VI.4. Members of the WOS all have a sense of justice. They are therefore not egoists. Since the conclusion of the congruence argument concerns members of the WOS, those arguments cannot be intended to show the egoist that it is good for him to be a just person (*TJ*, p. 567f/497f). It may be thought that, insofar as they judge from within the thin theory of the good, members of the WOS are acting like egoists or are judging as the egoist would. If this were so, then a successful argument for C_6 would imply that the egoist would judge that it is good to be just. But as we shall see much later, there is a great difference between the person who has a sense of justice but judges from within the thin theory, on the one hand, and the egoist, on the other. The difference, according to the Rawls of *TJ*, is that the person who has a sense of justice also has certain other-directed final ends such as friendship that she values even from within the thin theory. The egoist either lacks those ends or does not treat them as final.

§II.4: Congruence and Inherent Stability

I have not yet shown why the stability that results from establishing the *Congruence Conclusion* would be inherent stability, or that establishing *TJ's*

Nash Claim would show that the second condition of inherent stability is satisfied. If establishing these conclusions is to show inherent stability, then the fact that the balances of reasons referred to by *TJ's Nash Claim* tilt as they do must be brought about by the institutions of the WOS.

We shall see that Rawls tries to prove C_N, C_6, and C_C by asking whether they would hold of a typical member of the WOS. These are all conclusions about what place the sense of justice occupies among rational desires. If an argument about a typical or representative member of the WOS is to establish them, what must make that person typical or representative is the set of desires she has and the weights she attaches to them. But how, we might wonder, could *any* member of the WOS be typical or representative in this way? How could it be that members of the WOS are sufficiently similar in their desires, or in some relevant subset of desires, for any one person to typify them?

The Rawls of *TJ* answers that the institutions of the WOS shape the desires of those who live under them, encouraging sufficient convergence on the relevant desires and weights that C_N, C_6, and C_C are true. That, he thought, is one of the ways that justice as fairness, when institutionalized, generates its own support. It is because justice as fairness would encourage this convergence that its stability would be inherent rather than imposed. Rawls intimates that Hobbes was one of the first thinkers clearly to appreciate collective action problems and their implications for political philosophy.[24] Rawls thought that by distinguishing questions and viewpoints clearly, by identifying the best conception of a sense of justice, by making plausible assumptions about human psychology, by examining the educational effects of just institutions and—as we shall see—by drawing on Kant, he could solve the stability problems Hobbes had put on the agenda of political philosophy centuries before while avoiding Hobbes's own troubling conclusions.

Of course, the question of whether human beings are subject to coercive institutions because of their inherent tendencies to injustice is much older than Hobbes's problem. Different answers to *that* question reflect some of the deep differences between the Christianity of Augustine, who thought that political authority was needed because of human sinfulness, and of Thomas Aquinas, who denied that.[25] Showing that justice does not need to be imposed on us would shed light on that older questions. It would show, Rawls thought, that Augustine, Hobbes, and other "dark minds in Western thought" were wrong about political life, for it would show that a just society suits our nature.[26]

24. Rawls, "Sense of Justice," *Collected Papers*, p. 106; see also *Lectures on the History of Political Philosophy*, p. 79.

25. See my "Augustine and Aquinas on Original Sin and the Purposes of Political Authority," *Journal of the History of Philosophy* xxx (1992): pp. 353–76.

26. The phrase "dark minds" is taken from Rawls, *Lectures on the History of Political Philosophy*, p. 302 Rawls applies it to Augustine and Dostoevsky.

But despite the precision with which the Rawls set up the congruence problem, and the ingenuity with which he addressed it, Rawls came to recognize that the arguments for *TJ's Nash Claim*—roughly, the claim that each person would judge, from within the thin theory, that it is rational to remain the kind of person who answers justice with justice—relied on assumptions that were inconsistent with other parts of his theory. He saw that he had failed to show institutions of the WOS would stabilize themselves in the second way, and so had failed to show that justice as fairness would be inherently stable. The Rawls of *PL* spoke of showing "stability for the right reasons" rather than of showing inherent stability. Yet as we shall see, Rawls's underlying concern in the two treatments of stability was essentially the same. The inconsistency Rawls found in *TJ*'s attempt to show inherent stability prompted his turn to political liberalism, and the many changes he introduced between *TJ* and *PL*. Indeed, as we shall see in §IX.1, Rawls introduced the idea of an overlapping consensus to establish what his arguments for *TJ's Nash Claim* could not: that from an artificial but important point of view, each member of the WOS would judge that it is rational to preserve her desire to treat the principles of justice as supremely regulative when others do the same. Appreciating the great ambition of Rawls's attempt to show inherent stability, we can now see why correcting "an inconsistency of this kind should force such extensive revisions" (*PL*, p. xix). In Chapter III, we shall see what inconsistency Rawls found.

III

Ideals and Inconsistency

In part I of *TJ*, Rawls argued that his principles of justice would be agreed to in the OP. As we saw in Chapter II, he recognized that that agreement could be destabilized if members of the well-ordered society (WOS) believed it was in their interest to defect. After arguing for the principles, Rawls therefore needed to show how the agreement reached in the OP could be stabilized, so that the WOS would remain stably just. I argued that Rawls's account of stability was very ambitious. He hoped to show that justice as fairness, when institutionalized, would stabilize itself by generating its own supportive attitudes in those who live under it. This would show that the threat of collective action problems could be averted without appeal to a Hobbesian sovereign and that justice as fairness would be *inherently* stable.

In *TJ* and, as we shall see, in *PL* as well, Rawls treated the problem of stability in two parts. The conclusion of the first part, treated in *TJ*, chapter 8, is that members of the WOS would all develop a sense of justice. In the second part, Rawls argues—crudely put—that each member of the WOS would, on reflection, judge that it would be good for her to maintain her sense of justice. If the arguments are successful, then no one in the WOS ever has sufficient reason to defect from the agreement reached in the OP. The threat of collective action problems is averted and justice as fairness is shown to be stable. And if the institutions that implement justice as fairness are what bring it about that each person would judge that being just is good for him, then—at least according to *TJ*—the stability of justice as fairness is inherent.

Plans of life are drawn up and assessed from a viewpoint I called the "viewpoint of full deliberative rationality." This is the viewpoint members of

the WOS adopt when they reflect on whether being just is good for them and draw up their plans accordingly. A more precise way of saying that they would judge "on reflection" that being just is good for them would therefore be to say that they would judge that it is good for them from that point of view. When they adopt the viewpoint of full deliberative rationality, members of the WOS see that they have some reasons that tell against remaining just. The judgment that maintaining their sense of justice is good for them is a judgment about what they have reason to do on balance and how, on balance, they should draw up their plans. And so the second part of the argument for inherent stability is supposed to establish:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In *TJ*, the second part of the stability argument is found in Rawls's treatment of what he calls "congruence," and I called this conclusion the *Congruence Conclusion*.

Points of view are situations of choice and judgment. They are defined by the desires of, the information available to, and the rules of inference and decision used by, those who occupy them. When someone adopts the viewpoint of full deliberative rationality, she takes account all of her desires, including her desire to act from the principles of justice for their own sake. This may seem to limit the interest of C_C. Surely an argument for inherent stability would be more powerful if it showed that members of the WOS would judge their sense of justice to be good for them even when they reflected on their plans from a different point of view, one which left that desire out of account. Rawls therefore defines such a point of view, the point of view of what he calls the "thin theory of the good." In *TJ*'s treatment of congruence, he attempts to establish, not just C_C, but what I said is the stronger conclusion:

C_6: Each member of the WOS judges, *from within the thin theory of the good*, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In §II.3, we saw that a crucial step in the argument for C_6 and C_C is what I called *TJ's Nash Claim*:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans, *when the plans of others are similarly regulated*.

As we saw at the beginning of Chapter II, Rawls said he made the changes between *TJ* and *PL* "to resolve a serious problem internal to justice as fairness,

namely…the fact that the account of stability in part III of *Theory* is not consistent with the view as a whole" (*PL*, pp. xvii–xviii). I said that that inconsistency is found in *TJ*'s treatment of congruence. So to see why Rawls made the changes between *TJ* and *PL*, we need to see what inconsistency is involved in *TJ*'s arguments for C_N, C_6, and C_C.

§III.1: An Inconsistency in Justice as Fairness?

Here is what Rawls says about the internal problem in justice as fairness:

> the serious problem I have in mind concerns the unrealistic idea of a well-ordered society as it appears in *Theory*. An essential feature of a well-ordered society associated with justice as fairness is that all its citizens endorse this conception on the basis of what I now call a comprehensive philosophical doctrine. They accept, as rooted in this doctrine, its two principles of justice. Similarly, in the well-ordered society associated with utilitarianism citizens generally endorse that view as a comprehensive philosophical doctrine and they accept the principle of utility on that basis. Although the distinction between a political conception of justice and a comprehensive philosophical doctrine is not discussed in *Theory*, once the question is raised, it is clear, I think, that the text regards justice as fairness and utilitarianism as comprehensive, or partially comprehensive, doctrines. (*PL*, p. xviii)

The WOS is a society in which everyone accepts or endorses the same conception of justice. According to the "idea of a well-ordered society as it appears in *Theory*," then, the WOS is a society in which all members endorse justice as fairness. If the stability enjoyed by justice as fairness is inherent stability, then justice as fairness itself, when institutionalized, must bring about the endorsement.

I have said the argument that everyone in a WOS would endorse justice as fairness depends upon C_N, C_6, and C_C. If the institutions of the WOS are to bring it about that these conclusions are true, they must bring it about that each reasonable and rational person's balance of reasons tilts, and is seen to tilt, toward maintaining the desire to be just. People in the WOS may, of course, want things that they can only get by acting unjustly. They may, for example, regard certain things as good, such as extra money, that they can only get by acting that way. But Rawls thinks institutions can weaken the temptations to injustice by encouraging those who live under them to adopt certain views about goodness—more specifically, by bringing it about that members of the WOS see the expected payoffs of a life regulated by justice as better or more desirable than the expected payoff of the alternative kind of life. And so it is "on the basis of" that view of what is really good in life that members of the WOS make the judgments referred to by C_N, C_6, and C_C, and accept justice as fairness.

Note that if Rawls wants to show that justice as fairness would be inherently stable, then he has no alternative to showing that members of the WOS accept it on the basis of a view about what kind of life is worth living, since the only other way to secure stability is to impose it by means of Hobbesian sovereign or a dominant ideology. This is what Rawls means by saying here that "*an essential feature* of a well-ordered society associated with justice as fairness is that all its citizens endorse this conception on the basis of what I now call a comprehensive philosophical doctrine."

Rawls implies here that the problem he found with *TJ* is that "the text regards justice as fairness" itself as a "comprehensive or partially comprehensive, doctrine[]." On my reading, the problem Rawls is pointing to in the quoted passage is this. The stability argument in *TJ* had presupposed that all members of the WOS—"all of its citizens"—conclude that a life regulated by principles of justice is better than a life in which a desire to act from the principles is treated as one desire among others. They conclude that *because* the institutions under which they live have successfully encouraged them all to accept the same view of the good and *because* that view of the good—that "comprehensive philosophical doctrine"—is justice as fairness itself. This convergence on one view of the good marks a sharp contrast with *PL*. There, Rawls continued to maintain that everyone in the WOS would accept justice as fairness "on the basis of *a* comprehensive philosophical doctrine," but he denied that they all have to endorse it on the basis of the *same* comprehensive doctrine.

At first sight, this interpretation of the quoted passage is bound to seem puzzling. It seems to imply that there was a circularity in justice as fairness as Rawls originally developed it, and that the changes Rawls introduced in making the changes between *TJ* and *PL* were intended to eliminate the circle. The interpretation therefore clashes with the reasons Rawls gave for making the changes: to eliminate an inconsistency rather than a circularity. But my reading does not imply the presence of a circular argument in *TJ*. The appearance of a circle is simply due to the fact that, on my reading, the Rawls of *TJ* tried to show that justice as fairness is "self-reinforcing."[1]

To get a clearer idea of how justice as fairness reinforces itself, it will help to return to the mortarmen's dilemma introduced in §II.2. To show that a code of military honor would be stable, we need to show that the mortarmen value a life of honor above a life in which they decide whether to desert case-by-case. One way to show that would be to show that, because they have a sense of honor, they all want to live up to certain ideals. We might show that as part of being formed in a military ethos, they all come to aspire to ideals like camaraderie, loyalty, and brotherhood-in-arms, and that they all want to be the kind of person who does not let others down. These ideals require them to

1. Rawls, "Sense of Justice," *Collected Papers*, p. 106.

govern their lives by a code of honor. As they learn to aspire to these ideals, they become the kind of persons who discount whatever they expect to gain by desertion. They affirm and preserve their sense of honor, and the code of honor is stable.

Similarly, Rawls could try to show that justice as fairness would be stable by showing that, because they have a sense of justice, members of the WOS would all want to live up to certain ideals. As part of learning to be just citizens, they would all come to aspire to certain ways of conducting themselves and their relations with others. Living up to these ideals requires them to regulate their lives by the principles of justice. Rawls could also try to show that as they learn to value these ideals, they become the kind of persons who discount what they could gain by injustice. Discounting the payoffs of injustice, they would then judge that their balance of reasons tips toward remaining just. They would affirm and preserve their desire to regulate their lives by the principles, and justice as fairness would be stable.

This, I believe, is roughly Rawls's strategy. But the way I have described the strategy can suggest that Rawls begins with certain ideals whose realization is of prior or independent value, and that he treats the principles of justice as directives for realizing them. If this were right, then acquiring the desire to act from principles of justice would seem to require that members of the WOS all come to want lives in which those independently valuable ideals are realized. We could then see why Rawls says that all the members of the WOS "endorse [justice as fairness] on the basis of what I now call a comprehensive philosophical doctrine"—namely, the comprehensive philosophical doctrine which accounts for the value of realizing the ideals. But this is not how Rawls proceeds. While he does think that as members of the WOS acquire a sense of justice they all learn to value certain ideals, he does not claim that the value of realizing those ideals is given independently. Rather, Rawls accounts for the value of realizing those ideals from within justice as fairness itself. To see how he does this, we need to look into Rawls's claim that *TJ* treated justice as fairness itself as a "comprehensive philosophical doctrine." Only then will we be able to see what Rawls meant by saying that in *TJ*, he had assumed that members of the WOS would all endorse justice as fairness on the basis of justice as fairness itself.

§III.2: Ideals and Comprehensive Conceptions

Readers too often assume that they know what is meant by a comprehensive doctrine. A comprehensive doctrine is, they think, something like utilitarianism or Kantianism, Millian liberalism or Thomist Catholicism: a fairly systematic body of ethical thought that provides answers to the big questions of human life. But talk of "something like" is too vague. When Rawls speaks of a comprehensive doctrine, he means something fairly precise.

Rawls says that a moral conception is comprehensive

> when 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 informs our conduct, and in the limit our life as a whole. (*PL*, p. 13)

We may be tempted to treat Rawls's use of the word "ideal" in this passage as if it were casual, and the list of conceptions he says are included in comprehensive doctrine as a list generated more or less at random to convey a general idea of the sorts of conceptions a comprehensive doctrine includes. But to accede to this temptation would be a mistake. The word "ideal" is used to refer specifically to ideals included in *TJ*. And Rawls lists friendship, association, and personal character precisely because those are the ideals that the Rawls of *TJ* thought were included in justice as fairness. Thus, there is nothing casual about Rawls's word choice or random about his list. When Rawls says that *TJ* treats justice as fairness as a comprehensive doctrine, he has something specific in mind. And what he has in mind is not just specific, it is actually specified and in just the place we would expect it to be—in his definition of a comprehensive conception.

What is an *ideal* of personal character, friendship, or association? To answer this question, let's return to the distinction Rawls draws between concept and conception. I referred to that distinction in §I.5 when I asked whether a metaphysical conception of the person is expressed by:

> (1.1) We are by nature free and equal rational agents who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

We saw then that various conceptions of the person specify the concept of the person by providing accounts—full or partial—of the powers, interests, and desires that persons have as such, and the standards by which they are assessed. So the concept of the person can be partially specified by an account of the principles in accord with which persons act and reason, for example, or the ethical principles by which their actions are evaluated. What results from this partial specification is a partial conception of the person. (1.1) expresses such a conception. Other conceptions can be provided by further refining (1.1)—by, for example, further specifying "free," "equal," and "rational." Real people can conform to, or deviate from, such a conception of the person. Someone's life conforms to a given conception of the person when he lives up to the principles of right action, or when he acts from the interests the conception says he has in virtue of being a person. As a first approximation, I believe Rawls thinks an *ideal* of the person is a partial conception of the person that is such that someone's conforming to it, or living in a way that satisfies it, is good.[2]

2. That Rawls takes an ideal as a kind of conception is suggested by his seamless transition from one notion to the other at "A Kantian Conception of Equality," *Collected Papers*, pp, 254–67, pp. 254–55.

This characterization of ideals is compatible with many accounts of the goodness of living up to ideals. One could have an account according to which conforming to an ideal is instrumentally good because it brings about some further end or state of affairs. I shall leave this possibility aside. The ideals with which I think Rawls is concerned are ideals such that their goodness is realized *in the conforming*. That goodness could be of many kinds. It could be aesthetic, moral, or political, for example. Furthermore, there can be different accounts of the source of goodness. It could be maintained that the realization of some ideal is intrinsically good. As we shall see, the account of goodness as rationality that Rawls lays out in the chapter 7 of *TJ* provides an alternative. According to that account, the realization of an ideal is good because it is rational for members of the WOS to value its realization. Finally, whether someone who conforms to an ideal, and thereby realizes the corresponding value, actually experiences the realization of that value as good will depend upon, among other things, her beliefs, desires, and qualities of character. It will depend—to use a turn of phrase I employed in §I.6—on her formation, including her formation by her political culture and by the institutions under which she lives.

When Rawls says that *TJ* "regards justice as fairness . . . as [a] comprehensive, or partially comprehensive, doctrine[]," he means that *TJ* regards justice as fairness as "includ[ing]" ideals understood in this way—as conceptions the satisfaction of it is rational to value. These conceptions of conduct, friendship, and association, then, are the ideals that justice as fairness includes. When Rawls implies that justice as fairness "includes" these ideals, I think he has a number of things in mind.

> First, justice as fairness uses its own distinctive accounts of human interests and powers, together with the principles of justice, to specify partial conceptions of conduct, friendship and association. These conceptions are such that, as members of the WOS live up to them, singly or together, they realize what they regard as important values.
>
> Second, the source of these values is itself accounted for by justice as fairness and, more specifically, by "goodness as rationality" understood as including what Rawls calls the "full theory of the good."
>
> Finally, at least by the time Rawls wrote *PL*, he had come to think that the desire to live up to those conceptions or ideals is central to a sense of justice that is informed by justice as fairness. In §II.1, we saw that the stability of justice as fairness depends upon a "match between the right and the good." (*TJ*, p. 577/505) As we shall see, one reason there is match is that the sense of justice includes a desire to live up to ideals the realization of which members of the WOS all have reason to value.

To illustrate the first thing Rawls has in mind, let me take one of these ideals—the ideal of personal conduct—as an example. Recall that the conception of a free and equal rational person, referred to by (1.1), is arrived at by specifying our ordinary or workaday concept of the person. On my reading,

Rawls develops the ideal of personal conduct that justice as fairness includes by further specifying the conception referred to in (1.1), spelling out the notions of freedom, equality, and rationality and appealing to the principles of justice to do so. The notion of freedom in (1.1), for example, is specified in part by what Rawls calls "full autonomy." While there are other important elements to the ideal of personal conduct—including other and complementary ways in which freedom is spelled out[3]—I shall concentrate on full autonomy because it was the part of the ideal Rawls came to find controversial.

Full autonomy receives its most extensive treatment in the original *Dewey Lectures*. There, Rawls says that full autonomy is realized by members of the WOS in their daily lives by

> affirming the first principles that would be adopted in [the OP] and by publicly recognizing the way in which they would be agreed to, as well as by acting from these principles as their sense of justice dictates.[4]

These conditions could be read as saying that members of the WOS realize full autonomy only if they affirm and act from *whatever* principles would be chosen in the OP, regardless of their content, and publicly recognize the way they would be agreed to. One might be inclined to this reading of the condition if one read Rawls in the same way Rawls says Sidgwick read Kant—as thinking that full autonomy is realized by acting from *any* self-legislated principles at all (*TJ*, p. 254f./224).

I believe Rawls thinks there is *a* kind of freedom realized by acting from the principles that would be chosen in the OP, simply because they have been chosen there, and that that kind of freedom plays a role in one of his arguments for congruence. That kind of freedom is available because the choice of principles behind the veil of ignorance is free choice, choice uninfluenced by various natural contingencies. Let us call this kind of freedom *thin autonomy*.

Rawls is not committed to the view that someone would realize thin autonomy by acting from, for example, the principle of utility if it had been adopted in the OP instead of his own two principles. And so acknowledging that there is such a thing as thin autonomy is not a concession that there is something to Sidgwick's criticism after all. Rather, thin autonomy is the contribution that "the way [Rawls's principles] would be agreed to" makes to the freedom members of the WOS realize when they act on them. But the content of Rawls's principles also makes a contribution to the freedom they realize. For the content of the principles is such that when the basic structure satisfies them, the development and execution of plans of life is also free. In a just

3. At "Independence of Moral Theory," *Collected Papers*, p. 299, Rawls refers to "the ideal of autonomous persons who take responsibility for their fundamental aims over the span of a life"; he makes clear in the *Dewey Lectures* that persons who take responsibility for their ends are free, but that the freedom they realize is not full autonomy.

4. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 315.

society, people's plans are not formed or lived out in response to morally irrelevant contingencies, such as someone's social position or her winnings in the natural lottery. It would therefore be a mistake to suppose that Rawls equates thin autonomy with full autonomy. Rather, he thinks that members of the WOS realize full autonomy when they act from principles chosen in the OP *both* because those principles are chosen behind the veil of ignorance *and* because of their content.

It is because the conditions of the OP force choice of principles with this content that Rawls says the OP makes good on the defect Sidgwick found in Kant (*TJ*, p. 254/224). Rawls's reply to Sidgwick in *TJ* thus suggests that the notion of full autonomy is at work in that book. But full autonomy is much more explicitly developed in the original *Dewey Lectures*, and contrasted with other notions, than it is in *TJ*.[5] And so on my reading, the original *Deweys* spell out an important point that was made less clearly before: that the ideal of personal conduct that justice as fairness includes is, in part, the ideal of the fully autonomous person.

For reasons we shall see below, Rawls thinks that members of liberal democratic societies—including the WOS—want to live as free and equal persons. They want, that is, to live up to the conception of the person as expressed in (1.1). Rawls argues that the best way for them to live up to that conception is to live up to the ideal of the free person, and so to the ideal of the fully autonomous person, as that ideal is specified in the original *Dewey Lectures*. The conception of the person expressed in (1.1) is not, he thinks, well enough specified for citizens to know what it requires. It is by representing persons as free and equal in the OP, defending the principles and arguing that the fully autonomous person acts from the principles, that Rawls hopes to provide the necessary specificity and guidance. That is why he says in the *Dewey Lectures* that "the aim of political philosophy, when it presents itself in the public culture of a democratic society, is to articulate and make explicit those shared notions and principles thought to be already implicit in common sense[.]"[6]

The conception of a free and equal rational person referred to by (1.1) therefore stands, as it were, halfway between our ordinary concept of a person and the ideal of personal conduct that Rawls says justice as fairness included in *TJ* and in other writings before the political turn. I said in Chapter I that (1.1) does not refer to a metaphysical conception of the person. It does not specify our workaday concept of the person by drawing on claims in metaphysics or philosophy of mind. The ideals of justice as fairness are not specified by drawing on claims in metaphysics either. But the Rawls of *TJ did* specify these conceptions or ideals by drawing on ethical values: he drew upon the

5. In the *Deweys*, the contrast is with rational autonomy rather than with what I have called "thin autonomy." Rawls says that rational autonomy is realized by the parties in the OP. Thin autonomy is realized by members of the WOS.

6. Rawls, "Kantian Constructivism," *Collected Papers*, p. 306.

value of autonomy when he further specified the conception of (1.1) into an ideal of personal conduct.

Thus, I think the conception of a free and equal rational person referred to in (1.1) was always intended—even in *TJ*—to express a noncontroversial conception of the person that is neither metaphysical nor drawn from a distinctive ethical view. Even in *TJ*, Rawls thought that conception expressed a way in which members of liberal democratic societies—members of the WOS and us, Rawls's readers—normally think of themselves. It is because (1.1) expresses the way members of these societies normally think of themselves that Rawls begins there. One of the most significant changes between *TJ* and *PL* is thought to be what is sometimes called the "relativization" of justice as fairness: the claim that justice as fairness is intended specifically for liberal democratic societies, rather than for societies regardless of time and place. The relativization of justice as fairness in *PL* is thought to constitute a moral retrenchment.[7] If I am right about why Rawls starts with (1.1)—and if what I said in §I.6 about the dependence of this self-conception on the educative work of liberal institutions is right—then this interpretation is a serious misreading of *TJ*. An important part of it was "relativized" to liberal democracy all along.[8]

Unlike the conception of the person expressed by (1.1), the ideal of a fully autonomous person—as found in *TJ* and as more fully presented in the original *Deweys*—*is* what the later Rawls would come to regard as a controversial ethical ideal. To see this, we need only compare the way full autonomy is presented in the original *Deweys* with the way it is presented in *PL*. In the original *Deweys*, Rawls says that the value of full autonomy "is realized only by citizens of the well-ordered society in the course of their *daily* lives"[9]—a characterization that leaves out distinctions Rawls would later take pains to draw. Those distinctions are clearly at work in the corresponding passage in the revised *Deweys*, which are found in *PL*. There, in a section entitled "Full Autonomy: Political not Ethical," Rawls says that full autonomy is realized by "citizens of a well-ordered society in their *public* life" (*PL*, p. 77, emphasis added). I take the subtle change of wording to reflect Rawls's later realization that his earlier description of full autonomy at least suggested a value that is

7. For some of many examples, see the sources cited in Leif Wenar, "The Unity of Rawls's Thought," *Journal of Moral Philosophy* 1 (2004): pp. 265–75, notes 7, 8, 9, and 12.

8. Bernard Williams spoke for many readers when he said Rawls "has only more recently said emphatically that the elaborate reflections of *Theory of Justice* are reflections for particular time and apply to a particular political formation, the modern pluralist state." Bernard Williams, "The Liberalism of Fear," *In the Beginning Was the Deed* (Princeton, NJ: Princeton University Press, 2005), pp. 52–61, p. 53. Perhaps Rawls has only recently said it emphatically, but I think that—at least in retrospect—*TJ* contains clear indications of its intended readership.

9. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 315 (emphasis added).

ethical and not merely political, one that governs "our life as a whole" (*PL*, p. 13) and not merely "public life."

But when Rawls says that justice as fairness "includes" an ideal, he does not only mean that that ideal is part of the theoretical apparatus of justice as fairness. He also means that justice as fairness gives an account of the goodness of realizing that ideal. In chapter 7 of *TJ*, Rawls makes clear that in justice as fairness, something is good just in case it is rational to value it. If conforming to the ideal of personal conduct—and thereby realizing the ideal of personal autonomy—is good, that must be because it is rational for members of the WOS to value that form of freedom.

To see why it is, recall that someone realizes full autonomy when she affirms and acts from the principles that would be adopted in the OP, knowing how they would be agreed to there.[10] I believe Rawls thinks it is rational for members of the WOS to value full autonomy in part because they want to express their nature as free and equal rational beings and, knowing the content of the principles and why they would be adopted, they know that only acting from principles with that content expresses their nature. I shall explain and defend this claim more fully in Chapter VII. For now what matters is that Rawls appeals to what he calls "goodness as rationality" to explain the goodness of full autonomy.

It will be important later that the account of the goodness of full autonomy is part of what Rawls calls the "full theory of the good" rather than the "thin theory." The distinction between the two theories of goodness is easily misunderstood, and I want to take a moment to spell it out.

The full theory of the good is part of the more inclusive theory of goodness as rationality, and so something is good according to the full theory only if it has the properties it is rational to want in things of that kind (*TJ*, pp. 399/350–51). What distinguishes the full theory is that it explains the rationality of valuing something by appeal to the content of the principles. Thus, Rawls says: "the characteristic feature of this full theory . . . is that it takes the principles of justice as already secured, and then uses these principles in defining the other moral concepts in which the notion of goodness is involved" (*TJ*, p. 398/349). The value of others' propensity to abide by the principles of justice is a clear example. According to the theory of goodness as rationality, this is a propensity that is valuable because it is rational for each member of the WOS to want it in his fellow citizens. Its value must be accounted for by the full theory because the rationality of wanting one's fellow citizens to act from this propensity depends upon the content of the principles (see *TJ*, pp. 435–36/382–83). The value of full autonomy is another example. It is rational for members of the WOS to value full autonomy because by living autonomously, they can realize something else it is rational for them to want: the expression of their nature as free beings. Full autonomy is therefore valuable

10. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 315.

according to goodness as rationality. The value of full autonomy can be explained only by the full theory of the good because whether someone expresses her nature by acting from principles she would give herself, and acting from them for their own sake, depends upon the content of those principles.

The thin theory differs from the full theory in accounting for the rationality of valuing objects of desire without appeal to the principles' content. The clearest cases of things that are good as defined by the thin theory are the primary goods. The rationality of valuing the primary goods does not depend upon the content of the principles. Primary goods are simply instrumental goods access to which everyone needs to advance his plan, whatever it is. Moreover, the goodness of primary goods *cannot* presuppose the principles, since the argument that parties in the OP would choose the principles depends upon their wanting primary goods.

The example of the primary goods can, however, be misleading. As I mentioned in §II.3 and as we shall see again in §V.1, in laying out the congruence arguments, Rawls asks us to imagine someone who reasons from within—or, as he puts it in a crucial part of *TJ*, who "follow[s]" (*TJ*, p. 569/499)—the thin theory of the good. Some readers suppose that such a person follows the means-ends reasoning exemplified by parties in the OP. This is a serious mistake. What is characteristic of objects whose value is captured by the thin theory is *not* that they are instrumentally valuable or that they are objects it is rational to want whatever else one wants. It is that the rationality of valuing them does not depend upon the rationality or the goodness of acting from principles that have the content Rawls's principles do. Someone who follows the thin theory can therefore value a wide range of objects and can value them as ends. She can, for example, value various ends associated with her religion and value them as ends. What she cannot do is value some objects because those objects are rational to want given a desire to act from Rawls's principles for their own sake. We will miss the strategy of the congruence arguments if we misunderstand the contrast between the full and the thin theories, and take too restrictive a view of what a person who "follow[s] the thin theory" can value.

With the distinction between the two theories of the good in hand, we can see that when Rawls said *TJ* treats justice as fairness as a comprehensive doctrine, part of what he had in mind was that justice as fairness was presented as including the ethical ideal of full autonomy in both senses of "include." Full autonomy is defined within justice as fairness, and justice as fairness provides a theory of goodness that accounts for its value.

The ideal of personal conduct that justice as fairness includes is not just that of someone who is fully autonomous. It is an ideal of someone who wants to conform to the principles of justice so that she can conduct herself according to principles she can sincerely avow before everyone else in the WOS. Her plan of life exhibits certain important kinds of rational unity. She treats the persons and forms of life to which she is attached as the principles of right demand.

The ideal of association that justice as fairness includes is the ideal that gets the most extended explicit treatment in *TJ*: the social union of social unions. That ideal is described in *TJ*, section 79, where its value is accounted for within the full theory. Rawls can therefore claim that members of the WOS can value realizing the ideals of personal conduct and association without supposing that doing so has an intrinsic value given independent of justice as fairness.

Thus in *TJ* and the original *Deweys*, the ideals included in justice as fairness are ethical ideals. Their inclusion marks justice as fairness as a comprehensive doctrine. But it is important that Rawls distinguishes partially from fully comprehensive doctrines. A conception is partially comprehensive "when it comprises a number of, but by no means all, nonpolitical values and virtues and is rather loosely articulated" (*PL*, p. 13). A conception is fully comprehensive "if it covers all recognized values and virtues within one rather precisely articulated system" (*PL*, p. 13). Some interpreters seem to think Rawls wrote *PL* because he came to see that *TJ* treats justice as fairness as a fully comprehensive view of the human good that is thoroughly secular and individualist.[11] Other readers deny that *TJ* contains a comprehensive doctrine at all.[12] Both of these interpretations of *TJ* are mistaken. There is nothing like a "precisely articulated system" of value

11. This interpretation is not generally presented in any detail. It discerns a fully comprehensive doctrine that is secular less in particular passages and arguments than in the tenor of *TJ* as a whole, and asserts that the Rawls of *PL* recognized the fully comprehensive view that had been in *TJ* all along.

Some evidence of the prevalence of this interpretation can be found in readings that blur the distinction Rawls drew in *PL* between partially and fully comprehensive views. Thus, Russell Hittinger seems to read *TJ* as presupposing a fully comprehensive doctrine; see his review of *PL* in *The Review of Metaphysics* 47, 3 (1994): pp. 585–602. See also Sheldon Wolin's review of *PL* in *Political Theory* 24 (1996): pp. 97–129, p. 103. Wolin simply equates "comprehensive doctrine" with "fully comprehensive doctrine," and assumes that *TJ* treats justice as fairness as an instance of the latter.

More evidence can be found in readings according to which *PL* is an attempt to respond to concerns religious citizens would have had about justice as fairness as originally presented. This interpretation is suggested by passages in Stephen Holmes's review of *PL* "The Gate Keeper," *The New Republic*, October 11, 1993, pp. 39–47. For example, Holmes says at p. 44 that "the main objective of [Rawls's] new theory is to avoid a traditional liberal bias toward the views and values of secular intellectuals"; the suggestion seems to be that *TJ* showed such a bias, and that in doing so it was offensive to "people who do not happen to hold a consolationless creed." In fairness to Holmes, I grant that other passages in his review suggest a different interpretation. Holmes says, for example, that "nothing in *A Theory of Justice* itself suggested that a just society had to … demand unanimity about moral ideals" ("The Gatekeeper," p. 39). But if we read "demand" as "require," then this passage veers toward the second incorrect interpretation, the one defended by Brian Barry; see note 12.

12. Thus, Brian Barry says that *TJ* "does not include 'conceptions of what is of value in human life, as well as ideals of personal virtue and character.'" see his "Search for Stability," p. 878.

explicitly presented in *TJ*, and so *TJ* does not present justice as fairness as fully comprehensive. But since *TJ does* present justice as fairness as including the ideals of personal conduct, friendship, and association, it treats justice as fairness as *partially comprehensive*. When Rawls says, in the passage from *PL* that I quoted at the beginning of §III.1, that *TJ* regards justice as fairness as a "comprehensive philosophical doctrine," we must take him to mean it regards justice as fairness as a partially rather than a fully comprehensive doctrine.

Rawls also implies in that passage that *TJ* treats justice as fairness as a comprehensive doctrine that is *shared*. As we have seen, what makes justice as fairness a comprehensive conception is that it includes ethical ideals. So when the Rawls of *PL* implied that in *TJ*, he had supposed that all members of the WOS share a comprehensive conception of the good, what he meant was that in *TJ*, he had supposed that all members of the WOS value conformity with, or the realization of, those ideals. Conforming to those ideals is part of each person's conception of the good. And so he thought that in *TJ*, he had accepted:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

The ethical ideals of justice as fairness are the partially comprehensive conception of the good life that members of the WOS were presumed to share. When Rawls implies that "the idea of a well-ordered society as it appears in *Theory*" is "unrealistic" because *TJ* assumes that everyone in the WOS will share a comprehensive doctrine, it is that partially comprehensive doctrine that he had in mind.

The question of *how* members of the WOS come to share this comprehensive doctrine brings me to the third thing I said Rawls has in mind when he implied that in *TJ*, he treated justice as fairness as a view that included ethical ideals.

By the time Rawls wrote *PL*, he was willing to say quite explicitly that the desire to live up to certain ideals is central to each person's sense of justice. To put it another way, the sense of justice centrally includes desires that the Rawls of *PL* calls "conception-dependent" (*PL*, p. 84) and that I shall call "ideal-dependent." In *TJ*, Rawls had argued that members of the WOS all normally acquire a sense of justice. Looking back on that argument from the vantage point of *PL*, I believe he thought that he had placed ideal-dependent desires at the heart of a sense of justice in his early work as well. If my conjecture is right, then Rawls thought *TJ*'s argument that members of a WOS would all normally acquire a sense of justice was, in part, an argument that they would all normally acquire the desires to be fully autonomous and to live up to the other ideals of justice as fairness. Since Rawls had also argued in *TJ* that each person's good consists in the fulfillment of her rational desires, the satisfaction of these ideal-dependent desires belongs to each person's good. The process by which everyone in the WOS acquires a sense of justice would therefore account for the fact that the partial conception of the good referred to by C_3 is generally shared.

There may seem to be something anachronistic about the claim that in *TJ*, the sense of justice includes an ideal-dependent desire for full autonomy, since the ideal of full autonomy received its most elaborate development in lectures published a decade after *TJ*.

I believe there is something to this worry, since Rawls's thought about the content and acquisition of a sense of justice underwent significant development between *TJ* and *PL*. In responding to H. L. A. Hart's criticism of *TJ*'s argument for the basic liberties, for example, Rawls drew heavily on the conceptions of the person and of social cooperation that he had introduced in the *Deweys* (see *PL*, pp. 300ff). From that point on, those conceptions and the ideals associated with them are, I believe, much more prominent in justice as fairness than before. This gave Rawls's theory an even more pronouncedly Kantian flavor than it had in *TJ* and required him to supplement *TJ*'s account of moral development in ways he never fully acknowledged. And so I believe that when Rawls wrote the passages in *PL* in which he explained his political turn, he was reading some of his later views about ideal-dependent desires back into *TJ*.

But I also believe that—in this respect, at least—Rawls quite rightly thought of the original *Deweys* as clarifying and elaborating *TJ* rather than as adding totally new elements. We have already seen that the ideal of full autonomy, at least, is to be found in *TJ*, in Rawls's reply to Sidgwick; later in *TJ*, Rawls says that the strength of the sense of justice depends upon "*the attractiveness of its ideals*" (*TJ*, p. 501/438, emphasis added). The place in *TJ* where Rawls most explicitly anticipates his later position is in a contrast he draws between his own view and rational intuitionism. There he says that according to intuitionism, the desire to be just "resembles a preference for tea rather than coffee" (*TJ*, p. 478/418). The clear import of this remark is that if intuitionism is right, then there is no reason for us to give much weight to the desire to be just. By contrast, the Kantian Interpretation of justice as fairness shows that the sense of justice is a desire "to act in accordance with principles that express men's nature as free and equal rational beings" (*TJ*, p. 478/418). Since, as we shall see in §IV.1, members of the WOS all have a highest-order-desire to express their nature, "the sense of justice aims at their well-being" (*TJ*, p. 476/417). The connection between the sense of justice, the expression of our nature and our well-being therefore enables the Rawls of *TJ* to argue that there *is* a point or a "rational aim" to living justly after all (*TJ*, p. 476/417). That is an argument he badly wants to make, since it shows an advantage of justice as fairness over one of its competitors. Since members of the WOS can fully express their nature only by realizing full autonomy, the argument seems to require that, even in *TJ*, the sense of justice entails an ideal-dependent desire to be fully autonomous.

In §IX.2, I shall discuss how Rawls's treatment of the sense of justice developed between *TJ* and *PL*. For now, suffice it to say that the right explanation of the developments draws together the three things I said Rawls had in mind when he implied that justice as fairness, as laid out in *TJ*, "includes" ethical ideals. The first of these is that justice as fairness uses the principles of justice

to specify ethical conceptions or ideals—like that of the fully autonomous person—from more abstract concepts, by specifying the principles from which the fully autonomous person acts. The second is that it accounts for the value of realizing those ideals using the full theory of the good. The third is that members of the WOS acquire the desire to live up to those ethical ideals as part of acquiring the sense of justice.

In brief, what Rawls came more explicitly to realize was that the principles of justice can be used to specify ideals such as the ideal of the fully autonomous person. In the WOS, these ideals would be publicly known parts of the political culture. The ideals make vivid what it would be like for everyone to act from the principles. In particular, they show members of the WOS how they could best do something they naturally want to do: live as free and equal rational beings. Seeing this connection, in a just society in which the ideals are actually realized, increases the motivation to act from the principles. The motivation to live up to the ideals does not depend upon perceiving some value, consequent on realizing the ideal, that is prior to or independent of the value of acting on the principles. Rather, the value of living up to the ideals is accounted for by the goodness of acting from the principles. The motivation to act from them comes from seeing and understanding what a just life is like and how it answers to the human good, as spelled out by the full theory. Once the principles of justice have been chosen, their implementation and publicity enable Rawls to "bootstrap" his way to a heightened motivation to comply with them.

Thus, Rawls came to think that ideal-dependent desires belong to each person's sense of justice, and that the satisfaction of those desires belongs to everyone's good. This coincidence brings about the "match" between justice and goodness that stability requires. But as Rawls came to see more clearly that the sense of justice is ideal-dependent, he also came to think that *TJ*'s account of stability had really depended upon C_3. And he then came to realize that that account was—for that reason—unrealistic. To see this, we need to see why justice as fairness would be stable if C_3 were true.

§III.3: Endorsing on the Basis of Shared Ideals

The Rawls of *TJ* thought that endorsement of justice as fairness by members of the WOS depends upon their convergence on the ideals included in justice as fairness, the ideals referred to by C_3. This is what he had in mind when he implied, in the passage I quoted from *PL* at the beginning of §III.1, that he had assumed everyone endorses justice as fairness "on the basis of" the same comprehensive doctrine. To "endorse" justice as fairness is not just to acknowledge the validity and soundness of an argument for the principles of justice, such as the Pivotal Argument laid out in Chapter I. To endorse it is to give all-things-considered acceptance to its claim to regulate one's practical reasoning. And so endorsement of justice as fairness requires the

judgment that a sense of justice informed by justice as fairness belongs to one's good.

I have claimed the Rawls of *TJ* shows that everyone in the WOS reaches that judgment by establishing three important conclusions. One is *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans, when the plans of others are similarly regulated.

Another is:

> C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

The third is the Congruence Conclusion:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

If I am right that Rawls thought the endorsement of justice as fairness depended on C_3, then his argument for all three of these conclusions must depend on it. But how does it do so?

Rawls provides a promising clue in the first passage in which he ventured an explanation of his political turn. That passage is found in a long footnote at the end of "Political Not Metaphysical." There Rawls says that the account of stability in *TJ* treated of what he calls "the simplest case"

> where the public conception of justice is affirmed as in itself sufficient to express values that normally outweigh, given the political context of a constitutional regime, whatever values might oppose them[.][13]

We shall see in §VIII.5 why Rawls referred to this case as "the simplest." For now, I believe that the values to which Rawls is referring to in this passage include the values of realizing the ethical ideals of conduct, friendship, and association that justice as fairness includes. These ideals specify particular *forms* of conduct, friendship, and association. The ideal of conduct is an ideal of a particular form of conduct: fully autonomous conduct. The form of friendship is friendship founded on justice. We saw that someone realizes full

13. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 414, note 33.

autonomy only if the principles of justice regulate her plan of life. The ideal of full autonomy illustrates an important fact about all the ideals that justice as fairness includes. Someone can realize any of those ideals in the WOS only if she is a certain kind of person: a just person. Thus, the ideal of the particular form of friendship specified by justice as fairness is an ideal that we can realize only if we are just. So anyone who values realizing the ideals included in justice as fairness, and who grasps the theory of justice, as members of the WOS are assumed to do, will value being just. Since by C_3 everyone in the WOS is assumed to value the realization of those ideals, everyone has reason to be a just person.

But the passage just quoted says more than this. For I take Rawls to be saying that in *TJ*, he assumed that the value everyone in the WOS attached to the realization of those ideals was such that it "normally outweigh[s]" competing values. Whatever members of the WOS think they might gain by free-riding, by evading their taxes, or by acting contrary to their sense of justice in any other way is normally outweighed by the good of realizing the ideals of conduct, friendship, and association. So based on the balance of their reasons, they each would affirm that having and acting from a settled, supremely regulative disposition to be just is part of their good.

Because the reasons that tip the balance are connected with the value members of the WOS attach to realizing ethical ideals, Rawls thought that everyone in the WOS would affirm justice as fairness "on the basis of" those ideals. When Rawls says he assumed that the public conception of justice is "sufficient to express" the value of realizing those ideals, I take him to mean that in *TJ* he took the ideals to be part of a sense of justice informed by justice as fairness, that he took the full theory of the good as sufficient to account for those values, and that he assumed no other ethical conceptions were needed to supplement the account. Each member of the WOS took justice as fairness itself to provide sufficient reasons for realizing those ideals; those reasons do not need to be supplemented or explained by further reasons drawn from, for example, a religious view according to which those ideals are worth realizing. That is why Rawls implies—in the passage from *PL* that I said suggested a worrisome circularity—that members of the WOS affirm justice as fairness on the basis of justice as fairness *itself*.

To make this account less abstract, let us return to Joan, the member of the WOS whom I introduced in §II.2. If C_3 is right, Joan wants to be a fully autonomous person. She wants to act from the principles of justice, so that she conducts herself according to principles she can sincerely affirm before others. She wants to be a just friend and a just citizen, supporting the institutions that insure her liberties. And she wants to participate in a social union of social unions by upholding the principles that make it possible. The ideals of conduct, friendship, and association with which Joan wants to conform are, as we have seen, ethical *conceptions* of the person, of friendship, and of association.

Joan knows that she may sometimes want to act against these desires. She knows that she may sometimes want to cheat on her taxes, or may be tempted

to rely on political principles that cannot be justified to others. But she also thinks that, at least in the circumstances of the WOS, the values to be realized by satisfying her ideal-dependent desires outweigh what she could get by acceding to those contrary desires. Treating her friends justly, for example, and acting from principles she can affirm before others are enduring parts of her good that she values more highly than she values ill-gotten money. Thus, Joan accepts the principles of justice on the basis of ideals that justice as fairness itself includes, in the three senses of "include" that I discussed. Furthermore, Joan thinks the goods of being a just person are *themselves* sufficient to tip her balance of reasons toward satisfying her ideal-dependent desires and preserving her sense of justice. She may have further reasons to be a just person, beyond the goodness of realizing full autonomy and the other ideals. But because *TJ* treated of stability in what Rawls calls "the simplest case," the Rawls of *TJ* would have said that she does not need them.

On my reading, the Rawls of *TJ* thought that Joan is a typical member of the WOS. Like Joan, everyone in the WOS accepts justice as fairness as a partially comprehensive doctrine. Each of them wants to live up to its ideals—just as C_3 implies. In §II.4, I argued that what must make the typical member typical is her set of desires and the weights she attaches to them. This is exactly what makes Joan a typical or representative member of the WOS. Like all members of the WOS, she has the ideal-dependent desires that everyone in the WOS has if it is true that:

> C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

So Rawls thought that everyone in the WOS would regulate his plans by his sense of justice and would do so because of the value he attaches to the ideals of justice as fairness. And so he can move directly from C_3 to the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

A society in which C_C is true is in equilibrium. The ideal-dependent desires that tip each person's balance in favor of being just are enduring desires, so C_C is true every time members of the WOS reflect on their sense of justice and the equilibrium is stable.

In §II.4, I said that Rawls thinks convergence on a common set of desires and weights is encouraged by just institutions. In the last section, I said Rawls thought the institutions of a WOS would encourage convergence on the ideal-dependent desires implied by C_3 as part of encouraging a sense of justice. The Rawls of *TJ* thought that encouraging convergence on those desires is one of the ways that justice as fairness, when institutionalized, stabilizes itself. On my reading, Rawls made the transition to *PL* because *TJ*'s account of stability depends on C_3 and

because he came to see that C_3 is unrealistic. The "simplest case," in which the desires to live up to the ideals referred to by C_3 were sufficient to tip everyone's balance of reasons, was too simple. The ideal-dependent desires included in the sense of justice need the support of a wide variety of comprehensive doctrines.

There may, however, seem to be a number of difficulties with the reading I have just sketched.

Recall that in §II.1, when I said how ambitious a task Rawls had shouldered in trying to show the inherent stability of justice as fairness, I said that the task of showing its inherent stability was that of showing that justice as fairness would be stabilized by the autonomous activity of members of the WOS. The problem of congruence is thus, in effect, the problem of showing how institutions can bring it about that each member of the WOS sees that his balance of reasons tips toward living autonomously. Now that we have seen how Rawls's solution to that problem depends upon C_3, his solution may seem to be trivial. For his solution seems to be that institutions encourage members of the WOS to live autonomously by encouraging their ideal-dependent desires to live as autonomous people. And if the ideal-dependent desires referred to by C_3 really are parts of the sense of justice, then I seem to be saying that institutions encourage them to live justly by encouraging ideal-dependent desires to be just. The argument from C_3 may establish the *Congruence Conclusion*, but that conclusion does not seem to be very illuminating.

Moreover, I have not said anything about how, on my reading, Rawls moves from C_3 to C_N and C_6. And it may seem that I cannot, because Rawls cannot get from one to the other two. In the previous section, we saw that the value of living up to the ideals to which C_3 refers is given by the full theory of the good. If each member of the WOS judges that her balance of reasons tips toward remaining just, but she makes that judgment because she values living up to the ideals referred to by C_3, she is making a judgment about how her balance tips from within the full theory. But C_N and C_6 concern the way each person's balance of reasons tilts *as judged from within the thin theory*, for C_N says:

> C_N: Each member of the WOS judges, *from within the thin theory of the good*, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

C_6 is identical to C_N, except for the "when" phrase. While the argument from C_3 may establish C_C, it does not and cannot establish the other two conclusions on which I have said *TJ*'s argument for stability depends.

The force of this objection is apparent if we recall why Rawls introduces the point of view of the thin theory. That point of view is, I said, his refinement of what he referred to elsewhere as "the self-interested point of view" (*TJ*, p. 336/295). The Rawls of *TJ* wants to establish C_N and C_6 to show that temptations that arise from that latter point of view are weakened or removed and that the WOS is "as stable as one can hope for" (*TJ*, p. 399/350). Showing

that is the real purpose of Rawls's treatment of congruence. It is hard to see how an argument from C_3 can establish a conclusion about how someone's balance of reasons looks from the self-interested point of view, or from the point of view of the thin theory, since the desires ascribed in C_3 are desires for objects that value of which depends upon the content of the principles of justice. And so congruence arguments premised on C_3 seem to be beside the point Rawls was really trying to make.

Finally, even if the interpretation I have laid out can be supported by some of the remarks Rawls made about why he took the political turn, it is not at all clear that it can be grounded in *TJ*. Though I argued in §III.2 that the ideal of full autonomy can be found in *TJ*, I found only hints that the Rawls of *TJ* appealed to ideal-dependent desires. There does not seem to be any place in Rawls's treatment of congruence at which he appeals to C_3. Since that treatment is what I said Rawls came to find unsatisfactory, my explanation of the political turn does not seem to be very well grounded in the text.

The first two objections do not tell against my interpretation. As we shall see, Rawls agrees that, considered one way, the case for congruence *is* trivial (*TJ*, p. 569/498). The triviality alleged in the first objection exemplifies just the kind of triviality Rawls has in mind. And as I intimated in §II.3, it is precisely because the argument that moves directly to C_C from C_3 seems trivial, or at least too weak, that Rawls offers arguments from C_N and C_6 as well. The second of the two objections just raised shows that C_3 is not a premise of those arguments, but the objection does not show that it plays no role at all. As to the third objection, I have said that the Rawls of *PL* read C_3 back into *TJ*, and not that he relied on it explicitly. To answer that objection, it is enough to show where he might have read it in.

In the next section, I shall try to show that Rawls read C_3 into the way he set up the problem of congruence in *TJ*. What Rawls came to find unsatisfactory about *TJ*'s treatment of stability was that it depended a solution to that problem that did not work *and* that the problem itself was badly posed. It was badly posed because it rested on the implausible supposition that everyone in the WOS had the same partially comprehensive view, and that partially comprehensive view was justice as fairness itself—understood, according to §III.2, as including the ideals to which C_3 refers. Far-reaching changes in justice as fairness were called for because of the importance of the threat Rawls introduced congruence to avert, and because he came to think that the way he posed the problem of congruence was fundamentally misconceived.

§III.4: Congruence and C_3

We have seen that the problem of congruence arises if we imagine a typical member of the WOS like Joan asking herself whether her plans should make room for the desires associated with her sense of justice. As I said a moment

ago, Rawls concedes that taken one way, the case for congruence is trivial or obvious. Let's look at the text:

> Now on one interpretation the question [of whether congruence obtains] has an obvious answer. Supposing that someone has an effective sense of justice, he will then have a regulative desire to conform with the corresponding principles. The criteria of rational choice must take this desire into account. If a person wants with deliberative rationality to act from the standpoint of justice above all else, it is rational for him so to act. Therefore in this form the question is trivial: being the sorts of person they are, the members of the well-ordered society desire more than anything else to act justly and fulfilling this desire is part of their good. Once we acquire a sense of justice that is truly final and effective, as the precedence of justice requires, we are confirmed in a plan of life that, insofar as we are rational, leads us to preserve and encourage this sentiment. (*TJ*, pp. 569/498–99)

This passage indicates clearly what the conclusion of the congruence arguments is supposed to be: that it is "rational" for members of the WOS to endorse a plan of life that "leads [them] to confirm and encourage" the sense of justice as "truly final" or supremely regulative of their plans. Plans are drawn up in the viewpoint of full deliberative rationality. Members of the WOS treat their sense of justice as regulative when they treat the desire to act from the principles as regulative, so the passage confirms my claim that the conclusion to be reached can be expressed as what I have called the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

I take the main points of the passage to be the following. The treatment of moral learning laid out in *TJ*, chapter 8 showed that members of the WOS would normally develop an effective sense of justice. Rawls implies that the sense of justice is or includes a desire "to act from the standpoint of justice above all else"—by which he seems to mean that those who have a sense of justice want to act from that standpoint "more than [they want] anything [else]" (*TJ*, p. 569/498).[14] This is a very strong claim, and a stronger claim than Rawls needs to show that the question of congruence can have an obvious answer. All Rawls needs to claim—and all I shall take him to claim—is that those with a sense of justice want "to act from the standpoint of justice" "more than [they want] anything" that they could secure only by acting unjustly.

14. See also Rawls's remark at *TJ*, p. 477/418 that "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."

That weaker claim is enough to show that, having acquired a sense of justice, members of the WOS "will then have a regulative [and effective] desire to conform with the corresponding principles."

If this desire is taken into account by someone asking whether she should adopt a plan that will "lead [her] to preserve and encourage" her sense of justice, then the conclusion of the congruence arguments, which I have identified as C_C, follows immediately. As Rawls implies in the sentences immediately following the passage I just quoted, inherent stability is established (see *TJ*, p. 569/499). But the immediacy with which the conclusion follows shows, Rawls thinks, that the question of whether congruence obtains "has an obvious answer" and the argument for that answer is trivial. This problem arises because of the way that question is interpreted: as a question to be asked and answered from the viewpoint of full deliberative rationality. For in that viewpoint, all desires—including the desire "to act from the standpoint of justice above all else"—must be taken into account. The clear implication of the passage, then, is that the question of congruence must be asked and answered from a different point of view.

Rawls's dismissal of the quick argument for C_C is maddeningly brief. For one thing, he does not say what is wrong with giving a trivial argument for C_C. The problem with it is worth spelling out.

I have argued that the treatment of congruence is supposed to help show that agreement on the principles of justice would not be undermined by "the generalized prisoner's dilemma" (*TJ*, p. 577/503). More specifically, it is supposed to help show that each member of the WOS would try to act from and maintain her sense of justice in the face of temptations not to, temptations that arise from what Rawls refers to as the "self-interested point of view" (*TJ*, p. 336/295). As I shall explain in more detail later, the treatment of congruence is supposed to help show *that* by showing that each member of the WOS would judge, on reflection in the appropriate viewpoint, that a just life is a good one and that she is glad she has the desire to live such a life. Someone who asks herself seriously whether a just life is a good one is not going to put her doubts and questions to rest by noticing that she wants to be just. Noticing this desire in herself, she will ask whether she is glad she has it. The real problem with a trivial argument for the *Congruence. Conclusion* C_C is that it fails to solve the problem the treatment of congruence is supposed to address.

Moreover, the quoted passage suggests that C_C is an obvious answer to the question Joan has asked herself because that question is, roughly, "Is it rational for me to maintain my desire to be just, given that I want above all else is to act justly?". This may be correct, but it oversimplifies. There are a number of questions to which C_C is an obvious answer. If C_3 is true, and true because the ideal-dependent desires to which C_3 refers are part of a sense of justice, then another is "Is it rational for me to maintain my desire to be just, given that I want above all else to be fully autonomous?". Still another is "Is it rational for me to maintain my desire to be just, given that I want above all else to participate in a social union of social unions?". What Rawls later thought, I believe, is that in *TJ*'s treatment of congruence, he had put *all* these questions aside so that he could give a nontrivial argument for C_C.

To see that Rawls's own later reading of his earlier work is plausible, we need to see what such questions have in common.

All ask about whether it is rational to maintain a desire to be just, given some further rational desire. Moreover, the objects of those further desires are objects the rational desirability of which depends upon the rational desirability of acting from the principles of justice for their own sake. The dependence is obvious in the case a desire to act justly. We have already seen the dependence in the case of full autonomy. This dependence is the reason that the questions have obvious answers. Someone who asks whether it is rational to plan to preserve her sense of justice, given that she has an effective desire to be fully autonomous, asks a question which is no less trivial than someone who asks whether she should plan to preserve her sense of justice, given that she has a desire to be just.

We saw in §III.2 that the full theory of the good accounts for value by appealing to the goodness of satisfying the principles of justice. What is common to all the questions I said the Rawls of *TJ* put aside is that they all ask whether it is rational to maintain the sense of justice, given the desire for an object the value of which can only be given by the full theory. As Rawls implies, the treatment of congruence can show what it is supposed to show only if C_C is the answer to a very different kind of question. To pose that kind of question, Rawls had to suppose that Joan leaves out of account, not just the desire "to act from the standpoint of justice above all else" but all the ideal-dependent desires to which C_3 refers. That supposition is integral to the way what Rawls calls "the real problem of congruence" is set up (*TJ*, p. 569/499).

What does it mean to say that Joan leaves all such desires out of account? Suppose that Joan assumes a point of view in which the only value she attaches to things is the value she would attach to them if she did not care about being just *as such* or *under that description*, and in which she does not want for its own sake anything else the value of which is given by the full theory. In this point of view, she may still care about being just or about being fully autonomous, but if she does, it will not be because she is moved by considerations of justice or by the ideals of justice as fairness as final ends. It will be because being just or being fully autonomous or being a member of a social union of social unions serves other interests she has.

I shall say more about this point of view in subsequent chapters, especially in §V.1. Here I shall just say enough to introduce it and to convey some idea of how the congruence problem is set up.

The point of view I am now supposing that Joan adopts may seem to be the point of view of a selfish or self-interested person. And that might seem to be just the point of view from which the congruence question arises in its most helpful and illuminating form. For justice as fairness will be destabilized if members of the WOS come to think, or come to think that others think, the desire to be just costs them too much. If Joan adopted a selfish point of view and saw that she still had compelling reasons to affirm her sense of justice, that would put any doubts about the costliness of that sentiment to rest and—assuming Joan is typical—help to show inherent stability.

One problem with describing this new point of view as "the point of view of self-interest" is that while Joan may be tempted to cheat on her taxes to have more money for herself—as the mortarman may be tempted to desert his post simply to save his own life—norms of right can also be undermined by temptations that are not properly described as "selfish." Joan may be tempted to cheat on her taxes because she wants extra money to pass along to her children or to give to her favorite charity. These temptations show that the notion of the "self-interested point of view" needs refinement. What really threatens stability are not just temptations to act selfishly, but *any* temptations that arise within the point of view of someone who is not moved by considerations of justice as such, or by ideals and ends the values of which depend upon the good of justice. Once we see the diversity of temptations that need to be outweighed, the description of the relevant point of view as "self-interested" seems inappropriate. But it is—by construction—the point of view of someone "following the thin theory of the good" (*TJ*, pp. 569–70/499).

Much later, we will see that as Rawls came to appreciate the pluralism of a WOS, he recognized that the temptations that threaten the stability of justice as fairness might well include temptations to act against the demands of justice for political, religious, or philosophical reasons. He then recognized the importance of establishing, not a claim about how balances seem to tilt when judged from within the thin theory, but how members of the WOS think those balances tilt when they judge "by their comprehensive view" (*PL*, p. 392). At the time Rawls wrote *TJ*, however, this shift lay far in the future. In *TJ*, Rawls says that the "real problem of congruence" concerns the person who adopts the viewpoint of the thin theory (*TJ*, p. 569/499).

In §II.3, I implied that the Rawls of *TJ* answers that problem in stages. He argues first that Joan would find herself faced with payoffs like those shown in Table II.3, where the payoffs are valued according to the thin theory of the good and where A > B > D > C.

Table II.3

		Player 2
	Maintain regulative desire to act from the principles	Decide case-by-case
Maintain regulative desire to act from the principles	A, A	C, B
Player 1		
Decide case-by-case	B, C	D, D

This shows that, even from the point of view of the thin theory, Joan judges that "the plan of life which [is regulated by the desire to act from principles of justice] is [her] best reply to the similar plans of [her] associates" (*TJ*, p. 568/497). Since Joan is typical, this establishes what I called *TJ's Nash Claim*, a claim I expressed as:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans, when the plans of others are similarly regulated.

In §II.3 I also indicated how, given the special circumstances of the WOS, Rawls can move from this conclusion to:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

C_6 expresses a conclusion Rawls explicitly says he wants to reach, the conclusion that "it is rational for [each person], as defined by thin theory, to affirm his sense of justice" (*TJ*, p. 568/497). And we have seen how Rawls can move from that conclusion to the *Congruence Conclusion*, C_C.

In Chapters VI and VII, I shall show that the congruence arguments Rawls offers in *TJ* are meant to establish C_C by way of C_N and C_6. Thus as the congruence problem is set up in *TJ*, there are two routes to C_C. The route that depends upon Joan's valuing objects of desire according to the full theory of the good, and that moves directly from C_3 to the congruence conclusion, is trivial and Rawls gives it only passing attention. The route that goes by way of the thin theory is the more arduous and demanding. That is where the Rawls of *TJ* thought that the real work of establishing congruence needed to be done.

The congruence arguments that go by way of C_N and C_6 depend upon desires that are referred to by four conclusions for which I shall argue in Chapter IV:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

The objects of these desires are, as we shall see, objects the value of which is given by the thin theory.

I said that Rawls came to believe that the problem of establishing congruence was inadequately framed in *TJ* because he came to think that he had assumed C_3:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

in setting up the problem.

In Chapter VIII, I shall consider the possibility that Rawls thought he assumed, not C_3 but the weaker:

C_3*: All members of a WOS want to live up to the ideal of full autonomy.

I believe that what Rawls says about what makes a doctrine comprehensive tells in favor of the former, but even if he thought he had endorsed the latter, the explanation of the changes between *TJ* and *PL* remains basically the same. For once Rawls took full account of the fact of pluralism, he came to regard C_3—or C_3*—as unrealistic. And once he came to doubt C_3—or C_3*—he did not just provide different arguments for congruence. He rethought the way the problem of congruence was set up. He came to think that the real problem of congruence did *not* lie in showing that members of the WOS would "endorse" justice as fairness "on the basis of" the desires referred to by C_4a, C_4b, C_4c, and C_4d or "from within the thin theory of the good." The real problem, he came to think, lay in showing that they would endorse it on the basis of their "reasonable yet incompatible comprehensive doctrines" (*PL*, p. xviii).

Where in *TJ*'s set up of the congruence problem did the Rawls of *PL* think he had assumed C_3?

We saw that in *TJ*, Rawls said that one way of posing the problem of congruence was put aside as trivial and another was the more interesting. These conclusions depended upon the assumption that Joan is a typical member of the WOS, and—as we have seen—typical because of her desires and their weights. That is an assumption Rawls came to think he had made because he came to think he had assumed that everyone in the WOS endorsed the same partially comprehensive doctrine and that that partially comprehensive doctrine was justice as fairness. It is assumption he came to think he made because he had shown that everyone in the WOS has a sense of justice, and he came to think that a sense of justice includes the ideal-dependent desires referred to by C_3 or C_3*.

If it is possible that C_3 and C_3* are false and that Joan is not typical, then there are three ways of posing the problem of the relation of the right and the good, rather than the two that *TJ* distinguished. There is the first way, as posed in *TJ* and understood in light of the *Deweys*:

> Is it rational for Joan to maintain her sense of justice on the basis of her desires for objects valued according to the full theory, including the objects of her ideal-dependent desires?

There is the second way, also posed in *TJ*:

> Is it rational for Joan to maintain her sense of justice on the basis of the desires referred to by C_4a, C_4b, C_4c, and C_4d?

Then there is the third question, which must be posed if C_3 and C_3* are unrealistic:

> If some members of the WOS do not have the ideal-dependent desires implied by C_3 or C_3*, is it rational for them to maintain their sense of justice on the basis of the various comprehensive views of the good they *do* hold?

The Rawls of *TJ* thought an affirmative but obvious answer to the first question followed directly from C_3 or C_3*, once we see what the ideal-dependent desires to which they refer are desires for. He thought that the second question posed the "real problem of congruence" because it supposed that Joan leaves her ideal-dependent desires out of account. He thought that an affirmative and interesting answer to it could be defended. But he did not see the need to take up the third question. What he came to think is that he had missed the need to take it up because he had assumed that C_3 or C_3* is true.

It may seem obvious that Rawls needed an overlapping consensus to account for stability once it became clear that the third question had to be answered. But the need to introduce an overlapping consensus is *not* obvious, for nothing I have said so far rules out the possibility that Rawls could answer the third question by answering the second. Even if C_3 and C_3* are unrealistic and members of the WOS do *not* converge on a partially comprehensive doctrine, if they all have the desires referred to by C_4a, C_4b, C_4c, and C_4d then they could all affirm justice as fairness on the basis of those desires. In that case, appeal to an overlapping consensus would be unnecessary.

To see why Rawls cannot proceed this way, and why he needs the account of stability he offered in his later work, we need to see *why* he came to think that C_3 and C_3* are unrealistic. I have said that a full appreciation of pluralism led Rawls to doubt C_3 and C_3*, but that was a convenient shorthand. For, as I insisted in §III.2, Rawls sketches an argument for C_3—or C_3*—in the original *Dewey Lectures*. He came to think that C_3—and C_3*—were unrealistic because he saw the weaknesses in the argument he offered for them. According to that argument, the normal development of the ideal-dependent desires that C_3 implies depends upon the presence of desires that are not ideal-dependent. As we shall see in Chapter VIII, it depends upon the presence of the desires referred to by C_4a, C_4b, C_4c, and C_4d. More specifically, the Rawls of the original *Deweys* thought that the development of the ideal-dependent desires implied by C_3 depends upon members of the WOS seeing that treating the principles of justice as supremely regulative is the best or the only way for them to satisfy the desires referred to by C_4a, C_4b, C_4c, and C_4d. The development of the ideal-dependent desire to be a fully autonomous person, for example, depends upon their thinking of themselves as free and equal rational beings and upon their seeing that the only way for them to satisfy the desire to express their nature as such beings—the desire referred to by C_4a—is by treating the principles that way.

This brings us to Rawls's arguments for an affirmative answer to the second way of posing the congruence question. For as we shall see in Chapters VI and VII, Rawls argues that it is rational for members of the WOS to maintain their sense of justice on the basis of the desires referred to by C_4a, C_4b, C_4c, and C_4d by arguing that they can best or only satisfy those desires by being just persons. Thus, if members of the WOS are to develop ideal-dependent desires referred to by C_3 and C_3* they must, in effect, see Rawls's argument for an affirmative answer to the second congruence question. Those arguments therefore form a vital link in the argument for C_3. As we shall see in Chapter VIII, Rawls came to doubt those arguments. It is because he came to doubt those arguments that he came to doubt C_3 and C_3*.

And so, having come to doubt C_3 and C_3*, and having come to see the importance of the third question, Rawls could not then answer that question with the arguments he used to address the second one. Once the third question was posed, appeal to an overlapping consensus was necessary. That is why, when Rawls later explained the shortcomings of *TJ*, part III, he said:

> the account of the stability of justice as fairness was not extended, as it should have been, to the important case of overlapping consensus…; instead, this account was limited to the simplest case where the public conception of justice is affirmed as in itself sufficient to express values that normally outweigh, given the political context of a constitutional regime, whatever values might oppose them.[15]

§III.5: C_3 and Inconsistency

To understand why Rawls made the changes between *TJ* and *PL*, we need to see why he accepts the conclusions C_4a, C_4b, C_4c, and C_4d and why he thought members of the WOS would normally develop the desires to which they refer. That is the task of Chapter IV. We will then see how Rawls appeals to those conclusions in arguing for congruence and why he came to think those arguments fail.

At that point, we will be able fully to appreciate the inconsistency Rawls thought undermined *TJ*'s treatment of congruence. Briefly put, the inconsistency is this. The Rawls of *TJ* thought that the inherent stability of the WOS depended upon members of the WOS having desires to live up to ideals that are included in justice as fairness. In his writings before the political turn, Rawls thought that the objects of their ideal-dependent desires were the ethical ideals referred to by C_3. As I have said, Rawls thought that the institutions of the WOS would encourage convergence by fostering a sense of justice. That is how those institutions generate their own support and it is why the stability that results is *inherent* stability. In sum, the Rawls of *TJ* thought that the institutions of the WOS would bring about the truth of C_3. But as I suggested in §I.6, and as I shall explain much later, he came to realize that those institutions also encourage pluralism, and that as a consequence of pluralism, members of the WOS would be unlikely to converge on how best to satisfy the desires referred to by C_4a, C_4b, C_4c, and C_4d, and on the ideal-dependent desires referred to by C_3. So the institutions that were supposed to bring about the truth of C_3 would also bring it about that C_3 is likely to be false. It is because of this inconsistency that *TJ*'s argument for the inherent stability of justice as fairness failed. This failure led to Rawls's political turn and to his re-presentation of justice as fairness as a political, rather than a partially comprehensive, liberalism.

15. Rawls, "Political not Metaphysical," *Collected Papers*, p. 414, note 33.

IV

The Acquisition of Four Desires

At the beginning of Chapter II, we saw that Rawls says he made the changes between *TJ* and *PL* because of problems with *TJ*'s account of stability. I argued in that chapter that the stability in which the Rawls of *TJ* was interested was *inherent* stability. Rawls wanted to show that the principles of justice which are were adopted in the original position could stabilize themselves.

A crucial part of showing that justice as fairness would be inherently stable consists in showing that members of the well-ordered society (WOS) would acquire a sense of justice. Another crucial part is Rawls's argument for the congruence of the right and the good. In his treatment of congruence, Rawls tries to show when members of the WOS reflect on their desires and plans from the appropriate point of view, they would attach greater weight to the goods available when they regulate their lives by their sense of justice than they would to whatever they could gain by not doing so. And so in this part of his treatment of stability, Rawls defends what I have called the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In the "Introduction" to *PL*, Rawls implied that in *TJ* he had assumed that everyone in the WOS would support justice as fairness on the basis of the same comprehensive doctrine: justice as fairness itself. In Chapter III, I argued that when Rawls said he assumed members of the WOS shared a comprehensive doctrine, he meant they shared a *partially* comprehensive doctrine. They

all, he thought, wanted to live up to certain ethical ideals that justice as fairness includes. More precisely, had assumed that:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

When he said that his treatment of stability depended on the assumption that members of the WOS all have the same partially comprehensive doctrine, what he meant was that *TJ*'s arguments for C_C depended, in various ways, upon the supposition that C_3 is true.

Rawls was understandably concerned that the stability of justice as fairness would be threatened by the temptations to various kinds of self-interested behavior. To clinch the case for inherent stability, Rawls needed to show that members of the WOS would not accede to those temptations. This would be shown most compellingly if it could be shown that things they want *apart from the objects of the desires referred to by* C_3 would still incline them to be just. And so the most interesting and powerful part of Rawls's treatment of congruence is the part in which he tries to show that members of the WOS would judge that the goods of maintaining the sense of justice outweigh competing goods, even when they judge their balance of reasons from the self-interested point of view. Given the way I said the Rawls of *TJ* refines "the self-interested point of view," it is the part in which Rawls argues for the conclusion:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

Because of its interest, C_6 answers what Rawls calls "the real problem of congruence" (*TJ*, p. 569/497).

Rawls's argument for C_6 proceeds in two stages. In the first stage, he argues for what I called *TJ's Nash Claim*:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

In the second stage, he argues that each member of the WOS would have the assurance she needs that everyone else's plans are regulated by their sense of justice.

The congruence arguments Rawls offers in *TJ* are devoted to establishing *TJ's Nash Claim*, since Rawls seems to think that if he can establish C_N, C_6 and the *Congruence Conclusion* follow straightforwardly. Once he came to doubt C_3, Rawls realized that the task of establishing C_N, and of showing congruence, had to be framed differently than it had been in *TJ* and that other questions had to be confronted than those he had posed in that book. To answer those questions, he needed to introduce the idea of an overlapping consensus. Thus once

he came to doubt C_3, Rawls came to think, not just that one or two arguments for C_N failed, but that the problem of congruence as he had set it up in *TJ* was fundamentally misconceived. It was misconceived because it left some of the most important questions about congruence out of account. But Rawls *did* come to think that *TJ*'s arguments for C_N were unsuccessful. We cannot see why he came to doubt C_3, or why he came to doubt *TJ*'s set-up of the congruence problem, without seeing what difficulties he found in those arguments.

On my reading, Rawls's arguments for C_N depend upon the following four claims:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

In this chapter, I shall look at how Rawls establishes these four conclusions. In Chapters VI and VII, I shall show how Rawls draws on them to address "the real problem of congruence." In Chapter VIII, we shall see why Rawls's deepening appreciation of pluralism led him to doubt some of these claims and some of the premises he had used to establish them.

To see why Rawls accepted C_4a, C_4b, C_4c, and C_4d in the first place is to see why he thinks members of the WOS would normally acquire the desires to which they refer. Since these conclusions are established as part of Rawls's much more sweeping account of inherent stability, we need to see how he thinks the institutions of the WOS would encourage those desires. The account of how they do so depends upon what Rawls calls the "Aristotelian Principle," on what he calls Aristotelian Principle's "Companion Effect," and on the psychological laws that govern human moral development. I want to begin by looking at the Aristotelian Principle, since I think that the interpretation of it that is almost universally received overlooks something of importance.

§IV.1: Two Readings of the Aristotelian Principle

The Aristotelian Principle says

> 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*, p. 426/374)

As Rawls states the "Companion Effect", it says:

> As we witness the exercise of well-trained abilities by others, these displays are enjoyed by us and arouse a desire that we should be able to do the same things ourselves. We want to be like those persons who can exercise the abilities that we find latent in our nature. (*TJ*, pp. 428/375–76)

In this section, I want to clarify the Aristotelian Principle, since I think it can be misread in ways that lead readers to overlook parts of Rawls's account of moral development that are important for my purposes.

The Aristotelian Principle as Rawls states it is a conjunction. The real interest of the Principle is generally taken to lie in the second conjunct, which asserts that human beings enjoy more rather than less complex activities. Let us call this the *Second Conjunct Reading* of the Aristotelian Principle.

The pervasiveness of the *Second Conjunct Reading* is attested to by the way the Principle is quoted and cited.[1] The reading is encouraged by what Rawls himself says about the Principle immediately after introducing it. He writes:

> The intuitive idea here is that human beings take more pleasure in doing something as they become more proficient at it, and of two activities they do equally well, they prefer the one calling on a larger repertoire of more intricate and subtle discriminations.

Part of the appeal of the *Second Conjunct Reading* is no doubt due to the fact that the Aristotelian Principle is supposed to help "account[ ]for our considered judgments of value" (*TJ*, p. 432/379). Some of what needs to be accounted for is the value we attach to activities like the arts, demanding intellectual endeavors and the appreciation of beauty. The second conjunct seems to be the part of the Principle that does that work, helping us to understand why rational plans include such activities, and doing so without appeal to perfectionist principles or to Mill's distinction between higher and lower pleasures. Despite the power of the second conjunct to do this work, I believe a reading of the Principle that locates all of its philosophical interest there is mistaken.

To see the mistake, let's look at an important passage in which Rawls lists three points that the Aristotelian Principle "conveys." The first two of these points are:

> (1) that enjoyment and pleasure are not always by any means the result of returning to a healthy or normal state, or of making up deficiencies;

1. For several quite different examples, see Rainer Forst, *Contexts of Justice: Political Philosophy Beyond Liberalism and Communitarianism* (Berkeley, CA: University of California Press, 1994), p. 58, note 24 and accompanying text and G. R. Steele, "Understanding Economic Man: Psychology, Rationality and Values," *American Journal of Economics and Sociology* 63 (2004): pp. 1021–55, p. 1036. See also Margaret Moore, *Foundations of Liberalism* (New York: Oxford University Press, 1993), p. 58 and Henry Richardson, "John Rawls," *Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/rawls/

> rather many kinds of pleasure and enjoyment arise when we exercise our faculties; and (2) that the exercise of our natural powers is a leading human good (*TJ*, p. 426, note 20/374, note 20).

As we saw a moment ago, the first conjunct of the Aristotelian Principle says that "human beings enjoy the exercise of their realized capacities (their innate or trained abilities)." Despite the undoubted importance of the second conjunct of the Principle, I take it that the first conjunct rather than the second is the one that conveys points (1) and (2). Since the *Second Conjunct Reading* locates the philosophical interest in the second conjunct of the Principle, it largely ignores the first conjunct. Philosophers who interpret the Aristotelian Principle according to the *Second Conjunct Reading* therefore neglect these two points. Yet (1) is a point we would expect Rawls to make if he thinks—as I suggested in §III.2—that the good of living up to ideals is realized in the living rather than in the results produced. If we neglect (1) and (2), we overlook the fact that, in asserting the Aristotelian Principle, Rawls is asserting that human beings enjoy the exercise of our natural powers and experience the exercise of those powers as a good. As we shall see in Chapter V, by overlooking this fact, adherents of the *Second Conjunct Reading* mistake—or are strongly tempted to mistake—the role the Aristotelian Principle plays in the congruence arguments with which Rawls became dissatisfied. They are then led to read those arguments in the wrong way.

Thus I favor a *Two Conjunct Reading* of the Principle, which acknowledges the philosophical interest of both conjuncts. Indeed, I take the second conjunct, while important, to qualify the first. By expressing (1) and (2), the first conjunct asserts a source of enjoyment; the second conjunct qualifies this assertion by asserting a condition under which that enjoyment is heightened. Though I agree that both conjuncts are significant, for present purposes I want to stress the importance of the first. Because the *Second Conjunct Reading* of the Aristotelian Principle is so pervasive, questions about the truth of the Principle typically concern the truth of the second conjunct.[2] But what of the first?

I take it that Rawls does not mean that we enjoy every exercise of our rational capacities. The *ceteris paribus* clause with which the first conjunct begins is presumably supposed to rule out that interpretation. To see whether the first conjunct is true, let us return to the two points I said that that conjunct expresses: that many kinds of pleasure and enjoyment arise when we exercise our natural powers, and that the exercise of those powers is a leading human good. Let us take our natural faculties to include our faculties of practical and theoretical reason, and those physical powers that are under our voluntary control. It seems clear that some exercises of our natural powers so understood give rise to pleasure or satisfaction, and that some of these exercises can be experienced as very great goods. Some of the activities in which these powers are exercised are therefore highly valued ends and ends valued

2. See Moore, *Foundations*, p. 58.

because they are exercises of our natural capacities. If this is all Rawls means to assert by the first conjunct, then the first conjunct would clearly be true.

In fact I think Rawls means to assert more, and that there is a third point expressed by the first conjunct which Rawls does not mention. I think he also means to assert that (3) some exercises of our natural powers are experienced as good and are highly valued ends *because* the powers exercised are natural powers, because they are part of our nature. If this reading of the first conjunct is correct, then it adds something important to the usual understanding of the Aristotelian Principle. For according to the *Second Conjunct Reading*, certain exercises of our natural powers are part of our good because of their complexity. (3) reminds us that those exercises are part of our good because of their connection with our nature.

Rawls offers an evolutionary argument for the second conjunct of the Aristotelian Principle (*TJ*, pp. 431/378–9). Considerations drawn from evolution also support the first conjunct. The powers that are natural to us are the powers human beings need to exercise to navigate the world. If we are to live at all, we need to use the powers of locomotion and of reason. It would be surprising if evolution favored creatures who need to exercise such powers in order to live, but who were incapable of enjoying the exercise of those powers or who always found their exercise painful or burdensome. Moreover, if creatures of some kind have to exercise certain powers, then it seems that creatures of that kind who are so constituted that they have an incentive to exercise them—incentives in the form of satisfaction or enjoyment—would enjoy an evolutionary advantage, and so would survive the process of natural selection. Thus, evolutionary considerations make it plausible that we would find the exercise of our natural powers satisfying, and that we would find it satisfying precisely because those powers are natural to us. I shall therefore take it that what is expressed by the first conjunct of the Aristotelian Principle is true.

The first conjunct of the Aristotelian Principle does not imply that we have an inborn desire to exercise our natural powers, or that the desire to exercise them emerges under any conditions whatever. Rather, the Aristotelian Principle is compatible with the claim that a certain amount of cognitive and emotional development is required if we are to become persons who desire and enjoy the goods to which the Aristotelian Principle refers. In fact Rawls, like Aristotle, thinks not only that some moral learning is necessary so that the requisite desires and capacity for enjoyment emerge, but also that that learning takes place in a social setting. The social setting he presupposes is, of course, the WOS of justice as fairness. This will not be surprising if we recall why Rawls provides an account of moral learning. That account is part of the larger argument for the inherent stability of justice as fairness. As we saw in §II.1, that larger argument purports to show that the institutions of the WOS would generate moral support for the principles of justice. We would therefore expect the account of moral development to be an account of how we would develop if we lived under those institutions.

§IV.2: The Acquisition of Four Desires

Recall that according to C_4a, C_4b, C_4c, and C_4d:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

I now want to show how Rawls draws on both conjuncts of the Aristotelian Principle, and the Companion Effect of the Principle, to argue for C_4a, C_4b, C_4c, and C_4d.

The Desire to Express One's Nature

At a number of points, Rawls says or implies that members of the WOS want "to express their nature as free and equal rational beings" (*TJ*, p. 445/390). As we shall see, the fact that he thinks we have this desire is very important to the arguments of *TJ* with which he eventually became dissatisfied. Surprisingly, there is no one place in the text at which Rawls explains why he thinks that members of the WOS would have that desire. I do not think he simply assumes it, but his explanation must be pieced together from a number of places in *TJ*. To see why Rawls thought members of the WOS would normally develop that desire, we need to look more closely into what our nature *is* and what a desire to express our nature is a desire *for*.

In the original *Dewey Lectures* and later, Rawls describes members of the WOS as "free and equal moral persons *who are both reasonable and rational*."[3] If he continued to think that we have a desire to express our nature as persons—as I believe he did for at least some time after he introduced this vocabulary—then he presumably came to think that what we have a desire to express is our nature as free and equal, reasonable and rational persons. Looking into what the desire to express our nature is a desire for must, it would seem, take account of this development in Rawls's thought. But the reason I want to understand the desire to express one's nature is to understand what role the desire plays in arguments with which Rawls later became dissatisfied. Because those arguments are in *TJ*, I am trying to understand the desire to express

3. Rawls, "Kantian Constructivism in Moral Theory," p. 532 (emphasis added).

one's nature in the terms in which *TJ* discusses it, and so I shall not employ the later vocabulary.

This omission is not as serious as it might first seem. *TJ* may not speak explicitly of our desire to express our nature as free, equal, and reasonable, but it does speak of our "desire to express our nature as *moral* persons" and it does so at a critical point in the arguments that ultimately interest me (*TJ*, p. 574/503). I do not think that there is any great difference between expressing our nature as moral persons and expressing our nature as reasonable and rational ones, but I shall not pursue this matter here. Instead, I shall simply assume that the clarifications Rawls made by explaining the moral in terms of the reasonable and rational do not change the shape of the account that follows here.

The desire to express one's *nature* is not like a desire to express one's *self*. Someone wants to express himself when he is aware of some distinctive feature of himself—a trait, an opinion, or a taste, for example—and when he wants to make others aware of that feature and aware of it as his. The desire to express our nature differs on both counts from the desire to express oneself so understood. First, what someone wants to express when he wants to express his nature is not something that distinguishes him from others. What he wants to express is what he shares with other human beings: his nature. Second, a desire to communicate or to make others aware of his nature is not part of the desire.

What the desire to *express* one's nature does include, I think, is a desire to *realize* one's nature or to *exercise* one's natural powers. To see this, recall the *Two Conjunct Reading* of the Aristotelian Principle. The first conjunct of the Principle implies that we have the desire I have equated with the desire to express our nature—the desire to exercise our natural powers. To see *that*, recall the two points that that conjunct expresses:

> (1) that enjoyment and pleasure are not always by any means the result of returning to a healthy or normal state, or of making up deficiencies; rather many kinds of pleasure and enjoyment arise when we exercise our faculties; and (2) that the exercise of our natural powers is a leading human good (*TJ*, p. 426, note 20/374, note 20).

Among the satisfactions referred to in (1) is the satisfaction Rawls thinks we experience when we exercise the faculties that belong to our nature. Since he thinks we find the exercise of those faculties satisfying, it would be natural for him to say that we have a desire to exercise them. Rawls thinks our good consists in the satisfaction of our rational desires. The point asserted in (1) is the ground for ascribing that desire to us. If (1) is the ground on which Rawls supposes that we have a desire to exercise our natural powers, we can see why he asserts (2) and why he thinks that a principle which expresses (1) also expresses (2). If we equate the desire to exercise our natural powers with the desire to express our nature, we can see that the desire to express our nature is a desire Rawls thinks we have because (1) is true of us. Acting on that desire, by expressing our nature, belongs to our good because (2) is also true of us. Thus the Aristotelian Principle—in particular, the two points expressed by the

first conjunct of the Principle—explains why expressing our nature is a good. As if to confirm this, Rawls says a few pages later that "from the Aristotelian Principle it follows that th[e] expression of [our] nature is a fundamental element of [our] good" (*TJ*, p. 445/390). In Chapter VII, I shall provide a much more detailed argument that connects the Aristotelian Principle and the expression of our nature with important human goods.

The line of thought sketched in the last paragraph suggests that for Rawls, unlike some other thinkers in the history of philosophy—such as the Augustine of the *Confessions*[4]—the desire to express our nature is neither a desire to achieve an end-state in which we find contentment nor a desire to be united with an object of human love that we naturally find completely fulfilling. Rather, taken together with (2), (1) suggests that for Rawls, the object of the desire to express our nature is—so to speak—*adverbial*. He thinks this desire is a desire to act in a certain way. The action Rawls has in mind is not a one-off action, nor does Rawls think that we express or realize our nature only at defining moments of life which, because they are critical, reveal what we truly are. Rather, I suggest, Rawls thinks that the desire to express or realize our nature as free and equal rational beings is a desire to live our whole lives as such beings. It is a desire which, if fulfilled at all, is fulfilled continuously in our deliberation and action. Rawls implies at one point, that "a person realizes his true self by expressing it in his actions" (*TJ*, p. 255/224). His use of the plural "actions" is revealing for it confirms, not only that we realize our nature in activity, but also that our realization of it is ongoing rather than one-off.

How do we realize or express our nature as free and equal rational persons in our ongoing activity?

Rawls thinks that each person's good is specified by a rational plan of life, a plan for living that she continues to modify and act upon. The desire to live as free and equal rational beings is, I suggest, a desire to form and execute our plans of life freely, rationally, and in a way that befits the status each of us has as the equal of others. Thus the realization of our nature is not best thought of as one among a number of ends for which our plans make room. Rather it is a higher-order end that we attain by adopting, scheduling, and pursing our lower-order ends in a particular way. It is an end we realize through ongoing, higher-order exercises of practical reason.

I believe Rawls thinks that these exercises of practical reason are guided by conceptions of ourselves. Some of these conceptions are descriptions that we take to apply to us, such as "teacher," "parent," "Canadian," "musician," and so on. Others are descriptions that can be ascribed to us even if we do not take

4. See *Confessions* I.1.1, where Augustine says to God "Still he desires to praise thee, this man who is only a small part of thy creation. Thou hast prompted him, that he should delight to praise thee, for thou hast made us for thyself and restless is our heart until it comes to rest in thee." This is a point on which Rawls seems to have broken with Augustine quite early; see my review of Rawls, *A Brief Inquiry into the Meaning of Sin and Faith*.

them to apply to us in any straightforward sense of "take," such as "inquirer." These identities are often tacitly held. Even so, they can provide us with practical reasons, depending upon our attitudes toward them. They may be identities we want to live up to, or whose demands we want to satisfy. Or they may be identities we wish to disavow, to repudiate, or to put behind us. Either way, these identities can guide our conduct and our reflection on it. They can therefore guide our exercise of practical reason. Conceptions of ourselves that actually do guide us are what Christine Korsgaard calls "practical identities."[5] Since I want to reserve the phrase "practical identity" and its cognates for another use, I shall refer to these conceptions of ourselves using the less-elegant label "self-conception."

The examples of self-conceptions that I have given so far include nationality, occupation, and familial and vocational role. Some of these are ascriptive, and some derive from roles that are voluntarily assumed. But a conception of what we are by nature can also be practical if, for example, that conception is one we wish to live up to or expresses a way that we desire to live. And so the conception of ourselves as by nature free, equal, and rational moral persons who choose their own ends can be a self-conception, for we can want to live in a way that befits such persons. I shall refer to this self-conception as the *free-and-equal self-conception*. I believe Rawls thinks that the desire to express our nature as free, equal, and rational is a desire to live up to the *free-and-equal self-conception* when we frame and execute our plans.

In Chapter VII, I shall connect the conception of oneself as rational with the higher-order interest in living lives or executing plans that exhibit various kinds of rational unity. Now I want to say something about the conception of oneself as free.

The Rawls of *PL* says—much more frequently than the Rawls of *TJ* did—that members of the WOS conceive of themselves as free and equal. We may be tempted to think that Rawls attached greater importance to the distinctive way members of the WOS think of themselves as he came more explicitly to base his theory on ideas drawn from specifically democratic culture. But this is not so.

That the earlier Rawls also thought members of the WOS have such a self-conception is suggested by an essay he published soon after *TJ*, in which he says that "citizens are to view themselves as free and equal persons."[6] That

5. Christine Korsgaard, *The Sources of Normativity* (Cambridge, MA: Cambridge University Press, 1996), pp. 100ff.

6. John Rawls, "Some Reasons for the Maximin Criterion," *Collected Papers*, pp. 225–31, p. 230; the essay was published in 1974. See also John Rawls, "A Kantian Conception of Equality," published in 1975, at *Collected Papers*, pp. 254–66, p. 255. The fact that members of the WOS have the view of themselves discussed in the text helps to explain why their representatives in the OP must be guided by a view of *them*selves as free and equal; see "Some Reasons for the Maximin Criterion," p. 227.

self-conception is connected with a higher-order interest that Rawls wrote back into *TJ* when he revised the book in 1999: the "highest-order interest in how all [one's] other interests, including even [one's] fundamental ones, are shaped and regulated by social institutions" (*TJ*, rev.ed., p. 131). For in the revised edition, Rawls says that people who view themselves as free

> do not think of themselves as inevitably bound to, or identical with, the pursuit of any particular complex of fundamental interests that they may have at any given time … Rather, free persons conceive of themselves as beings who can revise and alter their final ends and who give first priority to preserving their liberty in these matters. (*TJ*, rev. ed., pp. 131–32)

Taken together, these remarks from the revised edition of *TJ* show that persons who think of themselves as free have an interest in how institutions affect their ability to pursue and revise their conceptions of the good. But this claim is not new to the revised edition. Rawls presupposes this interest in the original version of *TJ*.[7] The claim that citizens of the WOS have this interest also seems to be at work in one of Rawls's early treatments of liberty, where he argues that parties to the original position would want to protect citizens' ability freely to change or reject their religious faith.[8]

Thus, the desire to live up to the *free-and-equal self-conception* is a desire to live in a way that befits persons who are free in this way, who have an interest in preserving this freedom and who, in particular, have an interest in how social institutions affect their choice and their revision of ends. This higher-order end is part of what the desire to realize our nature is a desire for. If I can show that all members of the WOS normally acquire a desire for this end as part of their moral development, and that realization of this end is part of their good, then I will have shown C_4a. Why think that this is so?

It could be that when Rawls first wrote *TJ*, he thought the *free-and-equal self-conception* was a presupposition of rational agency. He might have thought that this is a self-conception every rational agent has—perhaps tacitly—and that its content and necessity can be discovered by philosophical reflection. But there is no evidence of this in the text. Instead, I believe the Rawls of *TJ* thought that members of liberal democratic societies, including the WOS, absorb the conception of themselves as free and equal persons from the political cultures of their societies, and that democratic culture would encourage their desire to live in ways that befit persons who think of themselves that way. I believe Rawls also thought that in the WOS, their acquisition of this self-conception is encouraged by seeing other members of the WOS live up to it (cf. *TJ*, p. 471/413).

7. See the closing sentences of *TJ*, §63 at pp. 415–16/365.

8. John Rawls, "Constitutional Liberty and the Concept of Justice," *Collected Papers*, pp. 73–95, p. 87. "Constitutional Liberty" was written in 1963.

Finally, recall that (1.1) says:

(1.1) We are by nature free and equal rational agents who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

In §I.6, I suggested that members of the WOS come to think of themselves as (1.1) says they are by seeing public institutions treat them as if (1.1) were true of them, and by publicly justifying their treatment of members by appeal to it. (1.1) expresses the *free-and-equal self-conception* as I have interpreted it. So the way members of the WOS are treated by their institutions encourages that self-conception as well. Rawls believed that members all want to live up to a view of themselves which is publicly encouraged by the liberal democratic culture of the WOS. Acquiring this view of themselves is part of what I called in §I.6 the "educative" or "formative" effect of publicity.

The desires referred to by C_4b, C_4c, and C_4d are, I believe, desires Rawls thought all persons normally acquire in the process of moral development, though the ties to which C_4c and C_4d may extend unusually widely in a WOS. But I believe Rawls thought that the *free-and-equal self-conception* and the desire referred to by C_4a—the desire to express one's nature as free and equal—typically depend upon liberal democratic institutions and liberal democratic political culture. Without the influence of liberal democratic institutions and liberal democratic thought, the self-conception and the desire would not be widespread. Since important arguments of *TJ* depend upon the claim that they *are* widespread—in a WOS and, as I shall suggest later, among Rawls's readers—it must be that at least the part of *TJ* that includes those arguments is successful only if the influence of liberal democracy is presupposed. We shall see that desire referred to by C_4a helps to stabilize justice as fairness. Since the widespread presence of that desire depends upon the work of liberal democratic institutions, encouraging the desire to express our nature is one of the ways that justice as fairness, when institutionalized, stabilizes itself. The argument for C_4a, and the later argument that builds upon it, are important pieces of *TJ*'s argument for inherent stability.

Since living in accord with the *free-and-equal self-conception* would be a complex exercise of our natural powers of practical reason, the Aristotelian Principle—even on the *Second Conjunct Reading*—goes some way toward showing that living up to it is part of the good of members of the WOS. These are significant conclusions, since they help to establish C_4a. Does the *Two Conjunct Reading* of the Principle shed any more light on why living in accord with the *free-and-equal self-conception* is part of citizens' good in the WOS?

It would seem not, for the Principle implies that we find it satisfying to act in some ways that are natural to us: to exercise our natural powers. It does not say anything about acting in ways that we *believe* are natural to us. But recall that on my reading, the first conjunct of the Aristotelian Principle says that the exercise of our natural powers can be satisfying, and part of our good, *because* those powers are part of our nature. That, I said, is the third point the first

conjunct of the Principle expresses. This connection between the naturalness of our powers and their contribution to our good is a connection on which we are capable of reflecting. We can ask what faculties are natural to us and which exercises of those powers comport with the kind of beings we are. Our desires to conduct ourselves in one way rather than another and the satisfaction we take in our activities are sensitive to these reflections. They are sensitive, that is, to what we come to believe about what we are and about how it is natural for us to act.

The Aristotelian Principle must be interpreted with this fact in mind. Earlier, I said the Aristotelian Principle suggests that we are liable to acquire desires to act in ways that are natural to us. If what I have just said is correct, then the Principle suggests that we are liable to acquire desires to act in ways that we *believe* are natural to us. It therefore helps to explain how members of the WOS acquire the desire to express their nature as free and equal rational persons. Coming to believe that the *free-and-equal self-conception* expresses a truth about their nature, they are liable to acquire the desire to live in accord with that conception of themselves. And if this is so, then the Principle also helps to explain why living in accord with that conception is part of their good. For their conception of the good is sensitive to what they believe about their nature. This explanation, unlike the one the Principle provides on the *Second Conjunct Reading*, does not depend only upon the complexity of living in accord with that conception. In §V.2, when I look at alternative explanations of the differences between *TJ* and *PL*, the difference between these two explanations will prove to be important.

The Desire to Avoid the Costs of Hypocrisy and Deception

Deceiving others imposes psychological costs on the deceiver (for reasons that will become clear in §IV.5, I leave out of account the costs of guilt and the pangs of self-reproach). What I have in mind is that the deceiver must calculate what to say and how to act, rather than speaking and acting spontaneously. Deception involves extra effort. Extra effort entails extra costs. Extra costs are undesirable. We shall see in Chapter V that, in a critical place, Rawls assumes that members of the WOS all want to avoid these costs. But I do not believe an argument for C_4b can be recovered from *TJ*. Instead, I conjecture that Rawls thinks it follows from what he takes to be the uncontroversial claim that *everyone*—whether or not she is a member of the WOS—wants to avoid these costs, other things being equal.

The Desire for Ties of Friendship

Rawls might treat the desire asserted by C_4c like the desire asserted by C_4b. That is, he might simply assume that human beings normally need and want friendship, and conclude that all members of the WOS need and want it. I think, though, that Rawls's treatment of moral development provides the materials for an interesting argument for C_4c.

In the course of their moral development, members of the WOS pass through three stages. The second of these is what Rawls calls the "morality of association." This is the morality developed in "the association of the school and the neighborhood, and also such short-term forms of cooperation . . . as games and play with peers" (*TJ*, pp. 467–68/409). The rules of the associations at work in this stage of morality are presumed to be just, and to be known as just. All their participants benefit from collective observance of the rules, and know that they do. And so participants in the various associations know that those who act to uphold the rules are acting in ways that benefit them. This recognition gives rise to "friendly feelings toward them, together with trust and confidence" (*TJ*, p. 470/411).

But a relation characterized by friendly feelings is not equivalent to friendship. Friendship is a relationship with moral component, including the liability to distinctively moral sentiments such as guilt. "Once these ties are established," Rawls says "a person tends to experience feelings of (association) guilt when he fails to do his part" (*TJ*, p. 470/412). Rawls's idea seems to be that at some point in our development, feelings of affection come to be accompanied by the desire to give something back to those who have benefited us and for whom we feel affection. This desire is the desire to move the relationship to a different footing. It is a desire to be friends—understood as a moral relationship—with those for whom one feels affection. This is something members of the WOS normally come to want. The liability to guilt when we fail to do our part is a natural part of this maturation.

Another way to put the point I am attributing to Rawls would be this. In the course of growing up, members of the WOS normally develop positive sentiments for those who benefit them. Actually living as friends with those who benefit them requires, among other things, that they be willing to reciprocate for benefits received, that they be liable to feelings of guilt when they do not, and that they want to protect the interests of those for whom they have affection (see *TJ*, p. 487/426). Without these other feelings and dispositions, they are not actually living as friends. Their affection, whatever it is, is not a "friendly feeling"—a feeling characteristic of friendship properly so called. But Rawls thinks that at some point in moral development, people naturally want to live as friends with those for whom they have those feelings. At a crucial point in the argument in which Rawls draws on C_4c, he assumes that "one needs the[ ] attachments" of friendship (*TJ*, p. 570/500). If what I have said here is right, then a better way for him to have put the point would have been to say that the desire for friendship is a desire members of the WOS normally acquire as they mature.

The conditions of reciprocity that naturally give rise to friendship are conditions that can prevail in a wide variety of associations. So a wide variety of associations are conducive to the development of friendship. Rawls says "we may suppose that there is a morality of association in which the members of society view one another as equals, as friends and associates, joined together in a system of cooperation known to be for the advantage of all and governed

by a common conception of justice" (*TJ*, p. 472/413). So the associations in which friendship develops can include the WOS itself. Rawls certainly does not mean that members of the WOS will become intimates. But he does think it natural that their relations will be characterized by a kind of friendship. So Rawls thinks that in a WOS, those who have passed through the second stage of moral development will have the desire described by C_4c and that their ties of friendship will extend quite widely.

C_4c is a conclusion about members of the WOS. The considerations that tell in favor of it are drawn from the discussion of moral development in the WOS. But, as we shall see in §IV.5, it is important that the connection between living as friends with someone and the desire to be fair to her does not depend upon any particular standard of fairness. No doubt there are limits to the standards of justice to which friends can try to conform. But the claim that various moral sentiments are constitutive of friendship does not depend upon defining fair treatment or reciprocity by the principles of justice or by any other principles that would be chosen in the OP.

If we are to see what is wrong with some alternative accounts of the differences between *TJ* and *PL*, we need to understand the place of the Aristotelian Principle and the Companion Effect in the acquisition of desires referred to by C_4a, C_4b, C_4c, and C_4d. With the *Two Conjunct Reading* of the Aristotelian Principle in hand, we can appeal to the first conjunct to help explain why friendship is desired and is experienced as a good. To say that two people are friends is not simply to say that they stand in a morally significant two-place relation. Friendship is, as Aristotle emphasized, an *activity* engaged in with others. To be a friend of another is *to live with him* in certain ways. I have used the rather awkward locution "living as friends" in restating Rawls's argument for C_4c to emphasize this fact. If the argument for C_4c that draws on the psychological laws is correct, then the good of friendship is a good that can be realized only in activity that is perceived by the participants to meet some standards of fairness. And if that argument is correct, then the activity of friendship is natural to us. We are, as Aristotle emphasized, naturally social. The first conjunct of the Aristotelian Principle reminds us that engagement in natural activities can be experienced as a great human good because those activities are natural.

Now recall that the Companion Effect to the Aristotelian Principle says:

> As we witness the exercise of well-trained abilities by others, these displays are enjoyed by us and arouse a desire that we should be able to do the same things ourselves. We want to be like those persons who can exercise the abilities that we find latent in our nature. (*TJ*, pp. 428/375–76)

The Effect, like the Aristotelian Principle itself, operates in the second stage of development and does important work. In acquiring the morality of a just association, Rawls says that members of the association acquire various ideals: they develop the desire to live up to the demands of the various roles within it.

For those who are already playing the roles well exhibit traits and excellences that members admire and—according to the Companion Effect of the Aristotelian Principle—desire to emulate. Summing up this line of thought, Rawls says "when the moral ideals belonging to various roles of a just association are lived up to with evident intention by attractive and admirable persons, these ideals are likely to be adopted by those who witness their realization" (*TJ*, pp. 471–72/413).

The morality of association is eventually followed by the third stage of moral development, the morality of principles. But Rawls says: "even though moral sentiments are in this sense independent from contingencies, our natural attachments to particular persons and groups still have an appropriate place" (*TJ*, p. 475/416). So the desire to protect the interests of persons and associations to whom one is attached remains even at the last stage of moral development.

The laws of psychological development are laws of reciprocity, but they are not—as it were—laws of mutual exchange. The morality of association is not one in which someone comes to care for the good of others in order to advance his own good. Rather, Rawls says that the laws of psychological development "characterize *transformations of our pattern of final ends* that arise from our recognizing the manner in which the institutions and actions of others affect our good" (*TJ*, p. 494/432). The good of persons and associations, the protection of their interests, and being a good friend or associate, all are included among the final ends of a member of the WOS. These are goods they come to want for their own sake, and not as means to some further end.

The Desire to Participate in Forms of Social Life that Call Forth Their Own and Others' Talents

To see what desire is being asserted in C_4d, and to see why all members of the WOS have that desire, we need to look in *TJ*, Chapter 9, to the section on social unions. That section falls into two parts. The first consists of the pages Rawls describes as the "preface" (*TJ*, p. 527/462). It lays the groundwork for the second part of the section, the discussion of "how the principles of justice are related to human sociability" (*TJ*, p. 527/462). For reasons I shall explain in §IV.5, I shall restrict myself to the preface.

The conclusion I impute to Rawls—namely:

> C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

does not imply that members of the WOS want to participate in *all* social forms that call forth their talents and those of others. Nor does it imply that, though they may want to participate in just some such social forms, they are indifferent about which ones they take part in. Rather, it means simply that everyone wants to participate in some such form or other, but it is compatible with their choosing favored forms for a variety of reasons.

The conclusion C_4d is not one Rawls explicitly defends. It is, however, a claim he builds upon in the second part of the section on social unions and one for which the preface furnishes an argument. Just what "participate" means will emerge as we proceed. A quick reading of the preface can suggest that the argument for C_4d goes as follows.

(4.1) "Rational plans of life normally provide for the development of at least some of a person's powers" (*TJ*, p. 523/458).

This seems to follow from the Aristotelian Principle or, as Rawls says, "the Aristotelian Principle points in this direction" (*TJ*, p. 523/458). Various human limitations cited by Rawls imply that:

(4.2) "everyone must select which of his abilities and possible interests he wishes to encourage; he must plan their training and exercise, and schedule their pursuit in an orderly way" (*TJ*, p. 523/459).

The training of our abilities and the scheduled pursuit of our interests normally require that we enter into associations with others. We cannot do these things on our own. This fact, together with (4.1) and (4.2), seem to imply that:

(4.3) Rational plans will normally include entering into associations with others.

Associations can, of course, assume many forms. We might wonder whether some forms of association are more likely than other forms to help each person advance the purposes mentioned in (4.2). It might seem that Rawls's analysis of games is supposed to help answer this question. (*TJ*, pp. 525ff./460ff.) For the game analogy seems to establish:

(4.4) If associations have the defining features of a social union, then others develop and exercise their talents in the association at the same time as we do.

This suggests that each person, interested in developing his own talents, can expect that he and others will find taking part in a social union worthwhile. But is there any reason to think that each will find social unions especially worthwhile—more worthwhile or enjoyable than other social forms?

The Companion Effect of the Aristotelian Principle says:

(4.5) "As we witness the exercise of well-trained abilities by others, these displays are enjoyed by us[.]" (*TJ*, pp. 428/375–6)

The Aristotelian Principle, plus (4.4) and (4.5), seem to single out a social union as a form of social life that members of the WOS will especially want to be part of, since by taking part in them, each—it may seem—can enjoy the development of his own and others' talents. And so it may seem that Rawls can infer the desired conclusion:

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

This reading of Rawls's argument fits with his remarks at the beginning of §79 about the kind of argument he intends to offer in that section: one that helps to support a "comprehensive" theory while starting from "simple and reasonable conditions that everyone or most everyone would grant." (*TJ*, p. 521/457) The argument depends upon weak and seemingly individualistic premises about what members of the WOS value—namely the Aristotelian Principle and the Companion Effect. It moves from these weak claims to a surprisingly strong conclusion, for it seems to show that Rawls can conclude from these premises that members of the WOS value participating in "forms of life" with the defining properties of social unions: "shared final ends and common activities valued for themselves" (*TJ*, p. 525/460).

Unfortunately, this reconstruction of the argument for C_4d faces a number of difficulties.

According to this reconstruction, the Companion Effect does considerable work in the argument. If the reconstruction is correct, then we would have to find a place in the text where Rawls asserts it. Rawls may seem to assert it at just the place we would expect it, for he writes:

> When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept. (*TJ*, p. 523/459)

But what is asserted here is not the Companion Effect *simpliciter*. It is the Companion Effect plus some important qualifications.

The first qualification asserts the importance of being "secure in the enjoyment of the exercise of [one's] own powers." I believe what Rawls means to stress here is the importance of knowing that one will be able to exercise and enjoy one's powers, and of the self-esteem that results from this knowledge.

Another qualification asserts the importance of shared aims. Since social unions have shared aims, I believe the intention of that qualifier is to assert a claim that Rawls will rely on later when he draws on C_4d. That is the claim that, because participants in social unions have common aims, social unions provide especially propitious conditions for their participants to do what the Companion Effect says we do: appreciate others' excellences. This important claim is not justified by the argument as I have reconstructed, nor is it appealed to. What the argument does appeal to—at step (4.4)—is the proposition that social unions provide especially propitious conditions for *others* to *develop* and *display* their excellences. This is, however, a very different claim than the one we would expect the argument to rely on in light of the qualification with which the Companion Effect is asserted. That qualification is not

primarily about those who are appreciated; it is about those who do the appreciating.[9]

Furthermore, the argument now under consideration does not establish that members of the WOS need to share the aims of those whose excellences they appreciate or that they need to participate in activities with them. At most, it seems to establish that members of WOS value the existence of social unions and want to be aware of them. For it would seem that we could appreciate the developed talents of others simply as passive spectators of their athletic activity or their musical performances. A reconstruction of the argument for C_4d that does not show why actually *participating* in the social union is good misses something important about the example of games, which Rawls seems to introduce to make a point about the value of actually taking part in a shared activity. A reading of the argument that leaves it unable to show the importance of participation is especially problematic because the treatment of social unions is ultimately supposed to show that members of the WOS value the goods of community (see *TJ*, p. 520/456). It is hard to see how the argument can show that if it does not establish that we need to participate in communities with others to enjoy what those unions make possible.

These difficulties suggest that the quick reading of the argument for C_4d is mistaken. To see how a more plausible reading would go, let us begin with the first difficulty faced by the initial reading—the difficulty in taking account of how the Companion Effect is qualified. Why is it that when people participate in a social union, they are especially well-disposed to appreciate others' exercises of their talents?

We are likely to miss the answer if we think of social unions as clubs or associations of the like-minded who share an end because they share a common interest. Thinking of social unions in this way can suggest that joining a club provides the occasion for witnessing what others do. Once we accept this suggestion, it is hard to see why actually participating in a social union—rather than being a passive member—is necessary, since passive membership might accord someone an adequate place for observation. The initial reading of the argument seems to convey this misleading suggestion between steps (4.2) and (4.3), where it says that developing our talents normally requires that we enter into associations with others.

9. When I introduced the *Two Conjunct Reading* of the Aristotelian Principle, I said that the first conjunct asserts an activity we enjoy—namely the exercise of our natural powers—and that the second conjunct qualifies that assertion by adding a condition under which that enjoyment is heightened. Interestingly, the Companion Effect as used in *TJ*, §79 has a parallel structure. The Effect itself asserts an activity we enjoy—namely, others' exercise of their natural powers—and the qualifier asserts a condition under which that enjoyment is especially likely to be available.

According to the right reading of the argument, developing our talents requires us to *associate* or *act* with others. So the right way to read the fourth step of the argument for C_4d is as referring, not to associations but to ways of associating or acting with others. This suggests that that step is not (4.4), but:

(4.4') If the activities in which we develop and exercise our talents have the defining features of a social union, then others develop and exercise their talents in the cooperative activity at the same time as we do.

This step is supported by Rawls's game example. I believe, though, that the example of games is supposed to do more than establish (4.4'). It also helps us see why we should accept the qualified version of the Companion Effect on which Rawls relies.

Suppose a game has the defining features of a social union: common activity valued for its own sake, plus a shared aim. This supposition imposes obvious constraints on the spirit in which players take part. It is the satisfaction of these constraints that distinguishes playing a game and participating in a social union. To participate requires that a player value the activity for its own sake and share the aim of the activity. Participation in this sense can itself be satisfying because developing modes of play consistent with the constraints of participation is itself the sort of complex activity to which the Aristotelian Principle refers. Moreover, it is participation so understood, rather than playing or watching passively, that makes the other goods of a social union available.

Now suppose that players value playing for its own sake and have as their shared aim executing a good play of the game. Then even if the excellence of others is responsible for their defeat, they will—insofar as they appreciate the fact that a game was well played—appreciate the excellent play of others. For the excellent play of all within the rules is what makes the play of the game a good one. And so the good play of the game is not something that is appreciated separately from, or in addition to, the excellent play of—say—the pitchers, the batters, and the fielders. Rather, appreciating the excellent play of each is part of appreciating the good play of the game. So to participate in a social union is to be "disposed to enjoy the perfections of others." What is it, exactly, that participants in a social union enjoy about the perfections of others?

If a game has accepted standards of excellent play, then there are accepted ways of using natural human powers—to throw, for example—well. Those who know the game will know those standards, and so will see excellent play as exemplary of the way they themselves might try to play. They will see excellent play as a realization of powers they themselves have. Thus, a game is an activity in which "different persons with similar or complementary capacities…cooperate so to speak in realizing their common or matching nature" (*TJ*, p. 523/459). Furthermore, I believe Rawls thinks there is special satisfaction in knowing that one's own play has helped to elicit the excellent play of others. This satisfaction is available only to those who contribute to or participate in the activity.

If these reflections on the game example are correct, then we can see why Rawls asserts the qualified form of the Companion Effect, which we can now treat as the fifth step in the argument:

(4.5') "When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept" (*TJ*, p. 523/459).[10]

In fact, if participants enjoy others' perfections as excellences of their shared nature which they helped to bring about, then perhaps each participant sees the excellences of others as in some way his own. If so, then (4.5') is less a qualified version of the Companion Effect than a version of the Aristotelian Principle itself. As we shall see in Chapter V, Rawls suggests as much himself when he draws on C_4d and the argument for it.

Be that as it may, the new argument for C_4d, unlike the initial one, shows why members of the WOS want to participate in social unions, instead of simply being aware of them as passive spectators. There are, Rawls thinks, many activities with the defining features of social unions. The game example is supposed to be generalizable, and so to show that we have reason to value them all. In the second half of the section of *TJ* on social unions, Rawls argues that the WOS is a particular kind of social union—it is a social union of social unions. He concludes that members of the WOS have reason to want to participate in it.

I have belabored Rawls's argument, and his game example in particular, to show how Rawls establishes points upon which he relies in subsequent arguments. Those are the arguments Rawls revised in making the transition from *TJ* to *PL*. Indeed, as we shall see in §§VIII.3 and VIII.5, the later rejection of (4.5') had significant consequences for justice as fairness and for Rawls's hopes for political philosophy. It is important to see why he accepted those points in the first place.

To see another point on which he relies here, consider the fact that, as Rawls notes, a good and fair play of the game is possible only if players take the rules of the game as regulative of their own play (*TJ*, p. 526/461). If they do

10. This passage might be interpreted simply as asserting that we enjoy the perfections of others at least when we develop and exercise our own powers. Joshua Cohen seems to interpret the passage at "Democratic Equality," pp. 748–49. But this reading is compatible with others' realizing themselves in activities in which I do not participate. This is, I think, too weak a reading of the qualification Rawls asserts to the Companion Effect in the quoted passage. On the stronger reading of the passage I have tried to defend here, the qualification asserts conditions that are especially conducive to the operation of the Companion Effect, because those conditions are especially conducive to my appreciation of the perfections of others. Those conditions are that I and they are engaged in the activity in which we realize ourselves, that we share the aims of that activity, and that we agree to norms which give those excellences a "place" in it.

not—if they cheat—then what results is no longer as good a play. Moreover, taking the rules as regulative is part and parcel of valuing the game in certain ways, for it implies that one values certain excellences of play above winning, and it is that way of valuing the game that disposes players to appreciate the good play of others. This conclusion is, Rawls thinks, applicable to all social unions. If we have good reason to want to associate with others in forms of social life that elicit everyone's talents, then we have reason to take its rules as regulative of our participation.

In the game example, the rules of the game satisfy two different descriptions: the description "the rules of baseball," for example, and the description "rules which are regulative of a social union." It is because the rules satisfy both descriptions that we can move—as in the previous paragraph—*from*

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

via the claim that a given game is such a social form *to* the conclusion that they have reason to take the rules of the game as regulative. The game example is supposed to be generalizable. The rules and norms governing each social union satisfy two different descriptions, one of which says they are rules or norms of the game in question and the other of which says they are regulative of the corresponding social union. Since the WOS is a social union, the rules that regulate it satisfy two descriptions as well. They are "principles of justice," "principles that would be acknowledged in the OP," and "principles which regulate a social union of social unions." The availability of this *diversity of descriptions* is, as we shall see, crucial to the arguments about which Rawls changed his mind in making the transition from *TJ* to *PL*. So seeing one of the points of game analogy is critical to understanding that transition.

§IV.3: Four Desires and Thin Reasons

I have now shown why Rawls accepts C_4a, C_4b, C_4c, and C_4d. But in what sense of "have" do members of the WOS *have* the desires to which these conclusions refer? In what sense of "have" for example, do members of the WOS *have* a desire to express their nature as free and equal persons?

It might seem that they do not have these desires at all, since these desires may not be necessary to provide either first-person or third-person explanations of their actions. I suggested in Chapter III that, at least by the time he published the original *Dewey Lectures*, Rawls accepted:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

And so he thought that members of the WOS would have ideal-dependent desires, such as the desire to be fully autonomous. The most accurate belief-desire explanation of their actions in daily life might therefore appeal to those

desires. Appeal to a desire implied by C_4a, such as the desire to live as a free being—where "free" is not given any further specification—might not, therefore, give the most fine-grained and accurate explanation of their actions. Furthermore, if members of the WOS have the conceptual resources of justice as fairness at their disposal, they might well explain their own actions using the conception of full autonomy rather than the cruder concept of freedom.

Even if the most accurate act-explanations do refer to ideal-dependent desires, a desire to be autonomous is a desire to be free—albeit in a particular way. It would not be inaccurate for us to say of persons who want to be fully autonomous that they want to be free and it would not be inaccurate for them to say it of themselves. Furthermore, there may well be times and circumstances in which explaining decisions by a desire to be free would be more felicitous than explaining them by a desire to be fully autonomous.

As I have already suggested and as we shall see in more detail later, ideal-dependent desires such as the desire to be fully autonomous are encouraged by the just institutions of the WOS. In the previous section, I said that those institutions encourage the desire to express our nature as free, while purposely avoiding the stronger claim that they encourage the desire to live with full autonomy. In fact, it seems likely that the institutions of the WOS will encourage both desires because they encourage the use of more and less general ethical concepts. Members of the WOS of *TJ* will absorb the general conceptual resources of liberal democratic thought. They will learn to think and describe themselves as free and equal rational beings, and as bearers of rights, and will want to live as such. They will also acquire the conceptual resources of justice as fairness. And so they will come to think of themselves as capable of full autonomy, will see full autonomy as an attractive kind of freedom, and will want to live as fully autonomous persons. Because the concepts of freedom and autonomy are importantly different, the desires to live freely and to live autonomously are different desires, and members of the WOS can be said to have both.

In this their situation is like our own, like the situation of Rawls's readers. Our political culture presents us with concepts of different levels of generality. We learn to think of ourselves and describe ourselves as free persons who are entitled to certain kinds of treatment. We also learn to use the more specific language of rights and liberties to describe our freedom and the kind of treatment that is due us. There is no difficulty in saying of us both that we want to live freely and that we want to exercise our rights and liberties.

I assume that what is true of the desires referred to by C_4a is true of the desires referred to by C_4b, C_4c, and C_4d as well. Members of the WOS have these desires as well as the ideal-dependent desires referred to by C_3. And so it would not be inaccurate to say that satisfying the desires referred to by C_4a, C_4b, C_4c, and C_4d is part of the good of members of the WOS. The good of satisfying those desires—like the good of satisfying the ideal-dependent desires referred to by C_3—is therefore a partial conception of the good shared by members of the WOS.

How the four desires are satisfied will be the subject of Chapter V. There we will see that they provide members of the WOS reasons to be just. These reasons differ from reasons connected with ideal-dependent desires in a way that will be important. It is therefore worth saying why the desires referred to by C_4a, C_4b, C_4c, and C_4d can be ascribed to members of the WOS. Why not simply describe them as having ideal-dependent desires? And why not simply say that the partial conception they share is that of satisfying those desires?

Ideal-dependent desires do a great deal of important political and philosophical work in Rawls's theory. If members of the WOS can most accurately be said to act from their ideal-dependent desires, then those desires do much to stabilize justice as fairness. Those desires therefore play a central role in one of Rawls's major, if overlooked, contributions to social theory: his explanation of how collectively rational norms of cooperation can avoid being destabilized by collective action problems, even in the absence of a Hobbesian sovereign. Those desires also play a central role one of his major, if underappreciated, contributions to moral theory: his argument that a just society suits our nature. If the arguments for those contributions are to be plausible, Rawls needs to say where ideal-dependent desires come from. If the stability enjoyed by justice as fairness is to be inherent stability, then those desires must be encouraged by the institutions of the WOS. So if his account of inherent stability is to be plausible, Rawls needs to say *how* just institutions would foster ideal-dependent desires.

As I indicated in §III.4 and as we shall see later, the Rawls of the original *Deweys* would have said that the development of these desires depends upon the prior presence of certain natural desires. For example, members of the WOS develop the aspiration to live as fully autonomous beings in part because, under the right social circumstances, we are the kind of beings who naturally want to live freely, in the less robust sense of "free" at work in C_4a. That natural propensity is shaped and educated by the public conception of justice in a WOS as part of our development of the sense of justice. One reason for ascribing the desires referred to by C_4a, C_4b, C_4c, and C_4d, and for looking at their origins, is to show that Rawls can provide the necessary account of where ideal-dependent desires come from.

Another reason stems from the fact that, as we saw in §III.2, the ideals referred to by C_3 are goods the value of which is given by the full theory of the good. That is why, as we saw in §III.4, an argument from C_3 to the *Congruence Conclusion*

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

is trivial and does not solve the problem the notion of congruence was introduced to address. To solve that problem, what has to be shown is a conclusion about what judgment members would make when they adopt a different point

of view, and when they ask whether their sense of justice is good for them on the supposition that the only values they have are those given by the thin theory of the good.

Now consider the objects of the desires referred to by C_4a, C_4b, C_4c, and C_4d. The values attached to these objects *can* be accounted for without appealing to the principles of justice. The value members of the WOS attach to expressing their nature as free and equal rational beings, for example, can be explained by the Aristotelian Principle. Members of the WOS naturally value friendship, and fairness to friends is part of friendship. But the value of friendship to those who naturally desire it does not depend upon taking the two principles as the standard of fairness. The value members of the WOS attach to participation in forms of social life that call forth talents depends upon the Aristotelian Principle and the Companion Effect. It does not depend upon the principles of justice. Indeed, I deliberately considered only the preface of *TJ*, §79 precisely because the argument beyond that point presupposes the principles and so belongs to the full theory.

Thus, the value members of the WOS attach to satisfying the desires referred to by C_4a, C_4b, C_4c, and C_4d is value that depends only upon the thin theory of the good. Someone who did not have a desire to be a just person as such or to be fully autonomous could still value the desire to act from the principles as part of his good, just as—to return to an example from previous chapters—a mortarman who did not have any desire to act from a code of honor or to be an honorable soldier could still want to act as the honorable soldier does because he wants the friendship of his comrades and finds hypocrisy in their presence too hard to sustain. Showing that members of the WOS all have the desires referred to by C_4a, C_4b, C_4c, and C_4d therefore shows that they share a conception of the good that is partial but *thin*.

As we shall see in Chapter V, the value of the ends referred to by C_4a, C_4b, C_4c, and C_4d provides members of the WOS reasons to maintain their desire to act from the principles of justice. Thus it is precisely because the value of ends that satisfy the four desires can be accounted for within the thin theory of the good, that they—and hence the shared thin conception of the good they define—provide reasons "which the thin theory of the good allows for maintaining one's sense of justice" (*TJ*, p. 572/501). They provide what I shall therefore call each person's "thin reasons" to be just. Thus, the second reason for insisting that members of the WOS have the desires referred to by C_4a, C_4b, C_4c, and C_4d is to show the thin reasons they have and the thin, partial conception of the good they share.

V

Thin Reasons to Be Just

In Chapters II and III, I gave some indication of what Rawls means by congruence and why he thinks he needs to establish it. In this chapter, I want to begin looking at Rawls's arguments for congruence with some care. Those arguments are found in §86, the penultimate section of *TJ*. Since I have said that it was Rawls's dissatisfaction with *TJ*'s treatment of congruence that led to the changes between *TJ* and *PL*, I need to show just how those arguments go if I am to explain the changes. Unfortunately, the arguments for congruence are not easy to make out. Part of the difficulty of making them out is that the arguments are presented very briefly, for reasons I shall mention in §V.3, and are not well situated in either their immediate or their larger context.

The immediate context of the congruence arguments,*TJ* chapter 9, can easily strike even the most sympathetic reader as a grab-bag affair. The chapter holds a number of arguments that can be the objects of some fascination when taken singly, but that do not obviously belong together. My own opinion is that chapter 9 is more disciplined than it appears to be and that the impression of incongruous juxtaposition is created by the absence of adequate transitional and explanatory remarks. None of the topics taken up in chapter 9 is superfluous. Indeed, some of the earlier sections of that chapter establish claims Rawls relies on to argue for congruence. I gave some indication of this in §IV.2, where I showed that in the preface to the section on social unions, Rawls argues for

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

As we shall see shortly, this claim serves as a premise in *TJ*, §86.

Since the focus of my attention will be §86, where the arguments for congruence are brought together, I shall not work systematically through all the sections of *TJ*, chapter 9 to demonstrate their importance. In Chapter VII, however, I shall try to show that Rawls's §§83–85—the sections on dominant ends, hedonism, and "the unity of the self"—although almost completely overlooked in the enormous secondary literature on Rawls, are critical to one of Rawls's congruence arguments. As we shall see, the arguments of these sections respond to Rawls's concern with the unity of practical reason—a unity that is threatened if an agreement reached in the OP is vulnerable to collective action problems, as I noted in §II.3. Grasping the connections between *TJ*, §86 and the sections that immediately precede it also shows the surprising connection that ties congruence to a much earlier part of *TJ*: "the formal constraints of the concept of right," laid out in *TJ*, §23.

The explication of §86 will also be crucial to what I hope will be a reignition of interest in part III of *TJ*. That part of the book is not read or taught as frequently as part I, and has attracted far less critical commentary. One reason for this, I suspect, is that readers have difficulty seeing how the arguments of part III fit together and contribute to the larger project of defending justice as fairness. Much of part III still needs to be integrated into a single, coherent reading. Unfortunately, I cannot provide such a reading here. But by showing how the sections on moral development and on social unions support C_4d, as well as

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

and by showing how the congruence arguments of §86 draw on those conclusions, we gain a new perspective on how those sections fit into a single line of thought that culminates in §86.

Furthermore, by seeing how the argument that depends upon C_4c also draws on remarks about the moral sentiments in the section on the morality of association, we can also see how precisely those remarks—which may seem to be interesting asides—fit into the larger sweep of argument. Seeing how the argument that depends upon C_4a also depends upon claims about the unity of the self shows the continuity of *TJ*, §86 with the sections that precede it. Another of Rawls's arguments, one we shall look at in Chapter VI, draws on what look like passing remarks about the love of mankind in §§72 and 73 of *TJ*. Rawls's arguments for congruence also depend, surprisingly, on the discussion of regret that seems to be introduced offhandedly in *TJ*, §64. The reading of the congruence arguments that I shall provide therefore shows that part III fits together much more closely than is sometimes thought.

§V.1: Setting up the Problem

What I have called the *Congruence Conclusion* says:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In §III.4, we saw that if the treatment of congruence is to solve the problem it was introduced to address, the argument for C_C must go by way of

> C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

And we have seen that the argument for C_6 goes by way of *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

The arguments for C_N in *TJ*, §86 lie at the heart of Rawls's treatment of congruence. Those arguments fall into two parts. In the first part, Rawls lays out the thin reasons members of the well-ordered society (WOS) have to maintain a sense of justice. In the second, he argues that those reasons are decisive—they tilt the balance of reasons in favor of maintaining the sense of justice as supremely regulative when others take theirs as supremely regulative as well. This establishes C_N. Given a solution to the *mutual assurance problem*, C_6 follows. In §§V.3ff, I shall discuss the first of these two parts. I shall discuss the second part of Rawls's argument—his argument for the decisiveness of the thin reasons to be just—in Chapters VI and VII. Before I look at the reasons to be just, I want to return to the way Rawls sets up the congruence problem in *TJ*, §86, since his setup of the problem can easily be misunderstood.

In §III.4, I quoted a passage from *TJ*, §86 in which Rawls dismisses the trivial argument for C_C. Immediately after that passage, Rawls writes:

> The real problem of congruence is 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. We should not rely on the doctrine of the purely conscientious act. Suppose, then, that 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 thin theory of the good would indeed confirm this sentiment as regulative of his plan of life. (*TJ*, pp. 569–70/499)

This is not an easy passage to interpret. Other readers have read it very differently than I do.[1] Since I first introduced the problem of congruence in §II.4, I have said that the person Rawls invites us to imagine when he sets up this problem is Joan, a typical member of the WOS. Since Joan is typical, she has a sense of justice, acquired according to the processes of moral development sketched in *TJ*, chapter 8. In light of the original *Dewey Lectures*, I assume she has ideal-dependent desires and desires for other objects the value of which is given by the full theory of the good. But as we saw in §III.4, if Joan adopts the viewpoint of full deliberative rationality and asks whether it is rational to maintain her sense of justice while taking these desires into account, the problem of congruence becomes trivial. To set up the "real problem of congruence," or the problem of congruence in its nontrivial form, I said Rawls invites us to imagine Joan taking up a different perspective on her desires.

I have referred to that perspective as the point of view adopted when someone reasons within the thin theory of the good. In §III.4, I gave some idea of what it would be like to reason from within that point of view. The passage I have just quoted fills in the details. In that point of view, Joan asks herself whether it would be rational for her to maintain her sense of justice as part of her good even if she did not desire objects—such as full autonomy—that *so described* are valued in light of the full theory. This does not imply that she abstracts away desires associated with her sense of justice, or pretends that she does not want their objects. Rather, she notices that those objects can be described in many ways. They satisfy a *diversity of descriptions*. Under some of those descriptions, the objects of desires associated with her sense of justice are valued as final ends and the value they have as such is accounted for only by the full theory. But under other descriptions, the same objects have values that can be accounted for entirely within the thin theory. What Joan asks herself, then, is whether it would be rational for her to treat the desires associated with her sense of justice as regulative of her plan *even if* she valued their satisfaction only to the extent that, by satisfying them, she attained objects she wants, the value of which is accounted for by the thin theory.

For example, if it is true that

> C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship and association included in justice as fairness.

then one of Joan's final ends is being fully autonomous. We saw in §III.2 that to act with full autonomy is to act from the principles of justice for their own

1. Brian Barry describes the passage as a "false start" and says that "nothing turns on" the prominent mention Rawls makes of the thin theory in the passage; see his "Search for Stability," pp. 885–86. For an interpretation of the passage that is much more sophisticated than Barry's but, I believe, quite different than mine, see Freeman, *Justice and the Social Contract*, p. 163.

sake. That is why we saw in §III.4 that if Joan asked herself whether it would be rational to maintain her desire to act from the principles as supremely regulative, taking account of her desire for full autonomy as such, the question would be trivial. But suppose that what Joan wants when she wants to be fully autonomous is also what she wants when she wants the object of the desire referred to by C_4a: to express her nature as a free and equal rational being. Suppose further that the objects of the two desires are the same. We saw in Chapter IV that the value of expressing one's nature can be accounted for within the thin theory. Joan can therefore begin to answer the congruence question in its nontrivial form by imagining that she values her desire to act from the principles—and hence her desire for full autonomy—just to the extent that being fully autonomous satisfies her desire to express her nature. She then asks whether the weight or the value she attaches to expressing her nature is sufficient to tip her balance of reasons in favor of affirming her sense of justice. This illustrates what I take Rawls to mean when he says that "the real problem of congruence is 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."

This way of posing the congruence problem depends upon the supposition that Joan's desire to act from the principles, her desire for full autonomy, and her desire to express her nature have the same object. The example therefore illustrates the importance of the fact that the objects of the various desires associated with a sense of justice satisfy a *diversity of descriptions*. But this fact is not one that can be taken for granted. According to the doctrine of the purely conscientious act defended by intuitionists like Ross, "the sense of right is a desire for a distinct (and unanalyzable) object" (*TJ*, p. 477/418). And so "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*, p. 477/418, emphasis added).

I cannot rehearse Rawls's very interesting arguments against this doctrine.[2] For present purposes, suffice it to say that one of the ways in which his contractualism differs from intuitionism is in "suppl[ying] other descriptions of what the sense of justice is a desire for." Those descriptions include "a desire to be fully autonomous" and "a desire to act from principles that would regulate a social union of social unions." As we shall see, the congruence arguments of *TJ* exploit this *diversity of descriptions* to show what reasons Joan has to affirm her sense of justice. Those reasons stem from the desires referred to by C_4a, C_4b, C_4c, and C_4d. Because the objects of those desires are objects the value of which is accounted for by the thin theory, this "connect[s] [the sense of justice] with reasons specified by the thin theory of the good." It is by showing that those reasons are decisive that Rawls establishes C_N and C_6. That is why he says that "we must use these [other descriptions] to show that a person

2. I say more about the argument in my "John Rawls and the Task of Political Philosophy," *The Review of Politics* 71 (2009): pp. 113–25.

following the thin theory of the good would indeed confirm this sentiment as regulative of his plan of life."

Properly interpreting the passage I quoted from §86 of *TJ* is clearly vital to understanding the problem of congruence, and I believe my reading fits it. I have stressed that the person Rawls invites us to imagine in this passage occupies the standpoint of the thin theory rather than the viewpoint of full deliberative rationality. This reading accommodates Rawls's reference to "a person following the thin theory." It also shows how he can answer the question I said his treatment of congruence has to answer if it is to help solve the stability problem: the question of whether the sense of justice would be judged to be good "from the standpoint of rational persons who have [it] when they assess their situation independently from the constraints of justice" (*TJ*, p. 399/350). On my reading, that question is answered by showing something about the relative weights of the reasons that members of the WOS take themselves to have when they judge from that standpoint. It is therefore answered by establishing conclusions, on my reading the conclusions C_N and C_6, which concern their balance of reasons. These conclusions can be established only if the person Rawls asks us to imagine is typical of the members of the WOS in relevant respects, for only if she is typical do conclusions about *her* balance of reasons establish conclusions about *each person's* balance.

§V.2: The Aristotelian Principle and the Argument for Congruence

In §IV.1, I contrasted two different readings of Rawls's Aristotelian Principle—what I called there the *Two Conjunct Reading* and the *Second Conjunct Reading*. According to the *Second Conjunct Reading*, the philosophical interest of the principle lies in its second conjunct, which implies that human beings enjoy complex activities. According to the *Two Conjunct Reading*, the interest of the Principle also lies in its assertions that pleasure and enjoyment can be found in the exercise of our natural capacities, that the exercise of those capacities can be a leading human good, and that it is a good *because* those capacities are natural ones. I argued then that the more novel reading, the *Two Conjunct Reading*, draws our attention to features of moral development that adherents of the *Second Conjunct Reading* might easily overlook. Now that we have turned to Rawls's congruence arguments, we can see another—and related—reason why the contrast between the two interpretations is of interest. The two readings of the Aristotelian Principle suggest different interpretations of the arguments for congruence. This difference can be illustrated by contrasting my interpretation with Samuel Freeman's.

When discussing the good of the WOS in *PL*, Rawls remarks that "the exercise of the two moral powers is experienced as a good. This is a consequence

of the moral psychology used in justice as fairness. . . . In *Theory* this psychology uses the so-called Aristotelian Principle" (*PL*, pp. 202–3). The remark occurs in the course of an argument that living in a just society is a good. The context of the remark, its mention of *Theory*, and the fact that the capacity for a sense of justice is one of the two moral powers, all lead Freeman to think that the remark provides a promising clue to the structure of the congruence arguments in *TJ*. Thus, Freeman says of this passage "This makes it seem as if the congruence argument involves a straightforward appeal to the Aristotelian Principle."[3] He continues:

> The idea here would be that the capacity for a sense of justice is among our higher capacities. It involves the ability to understand, apply, and act from requirements of justice. 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.[4]

Freeman calls this the "simplified argument from the Aristotelian Principle." He then introduces two very powerful objections to that argument. First, he notes, the argument does not show why it is rational for everyone in a WOS to develop *this* complex capacity rather than some other. Second, the simplified argument does not support the conclusion that it is rational to make the sense of justice "supremely regulative of *all* our pursuits."[5] He then says:

> The simplified argument from the Aristotelian Principle is not Rawls's argument for congruence. But it is extremely difficult to piece together what his argument is. The best way to uncover his argument is by seeing how he would respond to the two objections just stated.[6]

The argument to be uncovered in this way is what Freeman refers to as "the Kantian congruence argument." The differences between *TJ* and *PL* are to be explained, Freeman says, by Rawls's attempts to remedy the deficiencies he later found in this argument.

One indication that my reading departs from Freeman's is that we take Rawls to be arguing for different conclusions. According to Freeman, the conclusion of Rawls's congruence arguments is that members of the WOS would affirm their sense of justice as regulative of their plans of life when they take all their desires into account. It is therefore a claim about how each would treat the sense of justice from the viewpoint of full deliberative rationality. Since the sense of justice is a desire to act from the principles of justice, the conclusion for which Freeman takes Rawls to be arguing could be reexpressed as a variant of the *Congruence Conclusion* C_C, a variant which reads:

3. Freeman, *Justice and the Social Contract*, pp. 156–57.
4. Freeman, *Justice and the Social Contract*, p. 157.
5. Freeman, *Justice and the Social Contract*, p. 157.
6. Freeman, *Justice and the Social Contract*, p. 157.

C_C': Each member of the WOS judges, from the viewpoint of full deliberative rationality, that she should maintain her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

On my reading, the congruence arguments laid out in *TJ* section are supposed to establish C_N and hence C_6—conclusions from which C_C follows. Both of these conclusions—unlike C_C'—concern each person's balance of reasons, and only one concerns how the balance seems to tilt from the viewpoint of full deliberative rationality.

I read the conclusions as referring to balances of reasons because, as we shall see, that reading is a better fit with the text and strategy of Rawls's congruence arguments. I also read them this way because doing so makes explicit that members of the WOS are comparing the payoffs of two ways of responding to the decisions of others. It therefore draws attention to the game-theoretic concerns that I have said motivate Rawls's treatment of congruence.[7] Freeman's reading, by contrast, leaves those concerns out of account. These may seem like relatively minor differences. In fact I think that the difference between the conclusions we impute to Rawls stems from a very deep difference between our interpretations. On my reading, members of the WOS compare payoffs following the thin theory of the good. It is only by seeing how they do this, and what conclusions they reach, that we can appreciate a deep and subtle point of Rawls's that I shall stress in Chapters VI and VII. The sense of justice transforms someone who has it, so that she values things very differently than the unjust person does, even when she leaves her desire to be just out of account. The transformative effect of justice as fairness, when institutionalized, is what makes justice as fairness inherently stable.

Attending to Freeman's objections to the "simplified argument from the Aristotelian Principle" can certainly bring to light important features of the arguments Rawls offers for congruence. As we shall see, Freeman's interpretation of the Kantian congruence argument also highlights something important about the sense of justice: for the person who has affirmed her sense of justice as supremely regulative, weighing her sense of justice in the balance against other desires is simply out of the question. But I worry that the interpretation also misleads by ignoring the fact that the decision to affirm the sense of justice as supremely regulative is a decision that depends upon a claim about the balance of reasons, and by elevating the Kantian congruence argument at the expense of the other congruence arguments Rawls offers precisely because the Kantian argument seems *not* to depend upon Joan's balance of reasons.

7. For an example of an interpretation of congruence that goes wrong by ignoring Rawls's game-theoretic concerns, see Larry Krasnoff, "Consensus, Stability and Normativity in Rawls's *Political Liberalism*," *The Journal of Philosophy* 95, 6 (1998): pp. 269–92, p. 285.

I do not want to lay out and criticize Freeman's interpretation here. Instead, I simply want to draw attention to his reason for thinking that the passage from *PL* that refers to the Aristotelian Principle suggests what he calls the "simplified argument." He thinks it because he interprets the Principle as stating simply that the development and exercise of complex capacities is experienced as a good. Thus, one reason Freeman starts down the path that leads to his interpretation of congruence is that he reads Rawls's reference to the Aristotelian Principle as a reference to its second conjunct.

An adherent of the *Two Conjunct Reading* of the Principle will not be drawn to that path because she will take Rawls's reference to the Aristotelian Principle differently. She reads the Aristotelian Principle as asserting that activities in which we exercise our natural capacities can be enjoyable and can be important elements of the human good. She therefore recognizes that the Aristotelian Principle, together with the Companion Effect, postulates a tendency to value certain activities as ends. She may think there are some activities—such as friendship and association—that all members of the WOS normally come to value as a part of their maturation because of the way the Principle and its Effect affect moral development in a just society. Recalling the remark that "from the Aristotelian Principle it follows that th[e] expression of their nature [as free and equal rational persons] is an element of their good" (*TJ*, p. 445/390)—and understanding the expression of our nature adverbially, as I did in §IV.2—she may think the expression of our nature is such an activity. She may think that the good of these various activities is an important part of what makes the experience of living in a WOS good for its members. And so she will naturally be drawn, not to what Freeman calls the "simplified argument," but to an argument for the congruence of justice and goodness that appeals to the goodness of these activities. If the arguments of Chapter IV are right, she will be drawn to a reading of the congruence argument according to which the argument appeals to the goodness of the ends referred to by C_4a, C_4b, C_4c, and C_4d.

§V.3: Four Thin Reasons

I noted above that in the first part of Rawls's congruence argument, he lays out the reasons members of the WOS have to affirm a sense of justice. On my reading, they have those reasons because they all have the desires asserted in C_4a, C_4b, C_4c, and C_4d. I have said that Rawls focuses on the typical member of the WOS, Joan. If my reading is right, then we would expect that in laying out the reasons, Rawls would assume that Joan has those desires, and would offer arguments that connect those desires to the desire to be just. And if what I have said about the importance of the *diversity of descriptions* is right, then we would expect Rawls to forge those connections by appealing to alternative descriptions of what the desire to be just is a desire for.

This is exactly what Rawls does. He lists four reasons "which the thin theory of the good allows for maintaining one's sense of justice" (*TJ*, p. 572/501). He runs very quickly through the arguments that Joan has those reasons because he thinks cataloguing the reasons is a matter of "reviewing various points already made" (*TJ*, p. 570/499). In my introductory remarks to this chapter, I indicated that there is a great deal to be learned about *TJ*, and about Rawls's later dissatisfaction with its treatment of stability, from his treatment of congruence. I want to start making good on those promises by going beyond Rawls's summary expositions and spelling out clearly the four arguments he has in mind.[8]

The Desire to Avoid Psychological Costs

The argument for the first of the four reasons is the least interesting philosophically. But of the four arguments it—and perhaps the fourth—most clearly illustrates the strategy I have attributed to Rawls. Here is Rawls's argument as I read it:

Joan has a sense of justice and is considering whether to treat it as a sentiment against which she can act if it suits her. The sense of justice is a desire to regulate her conduct by the principles of justice. Since the principles are chosen in the OP, subject to the publicity condition, the principles are "public" in the WOS. That implies that everyone in the WOS accepts the principles and knows that everyone else accepts them. The principles therefore "characterize the commonly recognized moral convictions shared by members of a well-ordered society" (*TJ*, p. 570/499). This implication supplies the second description of what Joan's sense of justice is a desire for. Not only is it a desire to act from principles of justice, but it is a desire to act in accord with principles that would be among the shared convictions of a just society.

Because Joan knows that everyone else has an effective sense of justice, she knows that she lives in a society in which the principles of justice are among the shared convictions of the society in which she lives. She therefore knows that all others will do their parts and that they will expect her to do hers. So if decides to treat her sense of justice as a sentiment she can act against, she will still have to pretend that she, too, has an effective sense of justice and acts from "commonly recognized moral convictions." As Rawls puts it, "since the conception of justice is public, [s]he is debating whether to set out on a systematic course of deception and hypocrisy, professing without belief, as it suits [her] purpose, the accepted moral views" (*TJ*, p. 570/499).

This "deception and hypocrisy" will impose psychological costs. But it follows from C_4b that Joan has a desire to avoid such costs. The only—hence the best—way this desire can be satisfied by someone following the thin theory in a just society is for her *always* to comply with "commonly recognized moral convictions." Since the principles of justice are among the "commonly recog-

8. I have chosen not to reproduce Rawls's arguments *verbatim*. I shall count on the interested reader to compare my reconstruction of the arguments with the original texts.

nized moral convictions" of the WOS, Joan can satisfy the desire referred to by C_4b only if she *never* acts contrary to her sense of justice. Thus, the only—and hence the best—way Joan can satisfy the desire is by preserving and acting from her sense of justice. Joan therefore has as much reason to maintain her sense of justice as she has to avoid the costs of hypocrisy and deception. Moreover, since everyone else in the WOS is just, those costs are not offset by the knowledge that others are also free-riding. This, Rawls seems to conclude, means that the *net* cost of hypocrisy and deception is high. The reasons Joan has to maintain her sense of justice are correspondingly strong.

The argument from C_4b does not depend upon the claim that the desire to avoid hypocrisy and deception is the same desire as the desire to be just person, that the desire to avoid hypocrisy and deception is constituted by a sense of justice or that no one—regardless of circumstance—could have a desire to avoid hypocrisy and deception unless he had a sense of justice. Rather, it depends upon the weaker claim that in the special circumstances of a just society, in which everyone else is just, Joan can be sure of satisfying the desire to avoid hypocrisy and deception only by being just herself.

I find the last step of Rawls's argument doubtful, and so it seems doubtful that the argument shows Joan has *strong* reason to maintain her sense of justice. I shall revisit these doubts in §V.5. For my present purposes, the difficulty with the last step is less important than the fact that the rest of the argument shows Rawls proceeding as I have said he does:

- Rawls assumes that everyone has a desire for some end, the value of which can be given by the thin theory: in this case, the desire asserted by C_4b to avoid the psychological costs of hypocrisy and pretense.
- He draws on one of the specifics of the contract theory—here on publicity, which is one of the "formal constraints of the concept of right" (*TJ*, §23)—to furnish an alternative description of what Joan's sense of justice is a desire for.
- He shows that, given the alternative description and the special conditions of the WOS, Joan knows that the best—because the only—way for her to satisfy the desire she is assumed to have is by conforming to the commonly recognized morality of the WOS.
- Conforming to the commonly recognized morality of the WOS requires that Joan maintain her desire to act from the principles of justice, and to treat that desire as regulative, when others are assumed to do so. So Rawls infers that Joan has as weighty a reason to maintain that desire as she has to attain the end he assumed her to want—in this case, the end referred to by C_4b.

The Desire for Ties of Friendship

The argument Rawls gives for the second reason follows the same template. It also shows clearly how earlier sections of *TJ* are drawn on to establish congruence.

To set up the argument, we again have to bear in mind what Joan is considering: whether to treat her desire to act from the principles of justice as a sentiment against which she can act if it suits her. The first argument appealed to the fact that a sense of justice is a desire to act from the "commonly recognized moral convictions" of a WOS. The second argument draws on a quite different description—one derived, not from the publicity condition imposed in §23, but from *TJ*, §74 on "The Connection between Moral and Natural Attitudes." According to *that* description, the sense of justice is a sentiment connected with the natural attitudes. I laid out Rawls's argument for that connection in the third part of §III.3. The argument shows, Rawls thinks, that "wanting to be fair with our friends and wanting to give justice to those we care for is as much a part of these affections as the desire to be with them and to feel sad at their loss" (*TJ*, pp. 570/499–500). Friendship requires that we treat others in ways that we and they regard as just and fair. Of course, different people in different times and places have accepted different standards of justice. Friendship among the Athenians of Aristotle's time may have required that friends treat one another according to the standards of Aristotelian rather than Rawlsian justice. This possibility does not, however, affect the point which I have drawn from *TJ*, §74: *a* sense of justice—understood for the moment as a desire to act according to some mutually recognized standards of justice—is necessary if one is to have ties of friendship at all.

C_4c says that members of the WOS desire ties of friendship; the argument I am now considering is conditional on that conclusion. It, like the argument from C_4b, is also conditional on the special conditions of the WOS. Those conditions include the fact that Joan and everyone else in the WOS has a sense of Rawlsian justice, a desire to act from principles of justice that would be chosen in the OP.

It follows from the connection between justice and the natural attitudes that Joan could not have attachments of friendship if she acted contrary to her sense of justice *whenever* it seemed to suit her. If we "assum[e] . . . that [Joan] needs these attachments," she will want to be fair to, and to protect, her friends. So "the policy contemplated [by Joan] is presumably that of acting justly *only* toward those to whom [she] is bound by ties of affection and fellow feeling, and of respecting ways of life to which [she is] devoted" (*TJ*, p. 570/500, emphasis added). The problem with this policy is that, as we saw in §IV.2, "in a well-ordered society," other people treat Joan justly and so "these bonds extend rather widely and include ties to institutional forms" (*TJ*, pp. 570–71/500). So Joan will care about being just toward, and will care about protecting, a large number of people, associations, and institutions. She will not want *them* hurt by her injustice. But she "cannot in general select who is to be injured by [her] unfairness" and passing along unjust gains and savings to those one cares about "becomes a dubious and involved affair" (*TJ*, p. 571/500). So the "natural and simple way [to protect] the institutions and persons [she] care[s] for" is to answer justice with justice. (*TJ*, p. 571/500) If natural and simple is best, then being

a just person when others are just will be the *best* way for Joan to protect the institutions and persons she cares for.

The desire for friendship referred to by C_4c, like the desire to avoid hypocrisy and deception referred to by C_4b, is not the same desire as the desire to act from Rawls's principles, nor is the desire to act from the principles constitutive of the desire for friendship, considered only as such. But given both the connection between moral and natural attitudes and the special conditions of the WOS, the desire to act from Rawls's principles is a desire to honor the demands of justice needed to sustain ties of friendship in a WOS. The desire for those ties and the associated desire to protect persons and institutions she cares for therefore give Joan reasons to preserve her sense of justice. And so Rawls concludes that "in a well-ordered society where effective bonds are extensive both to persons and to social forms, and we cannot select who is to lose by our defections, there are strong grounds for preserving our sense of justice." (*TJ*, p. 571/500)

As with the argument for the first reason, so with the argument for the second, we may doubt that Rawls has actually shown that Joan has *strong* reasons to preserve her desire to act from the principles of justice. The amount that Joan withholds from her tax payments, for example, may not be large in absolute terms. She may be able to convince herself that the damage she inflicts on others by cheating is small enough, and the burden spread widely enough, that no one she cares about is likely to be hurt much at all. Another question about the argument arises once we consider the possibility that Joan cares far more for some people than others. If she can benefit her family greatly by cheating on her taxes or by conniving to get a highway or a dump located near someone else's house, she may consider that that gain outweighs the cost to her friends or to others to whom she has more tenuous connections.

Rawls ultimately wants to show that it is rational for members of the WOS to preserve their desire to regulate their plans by norms of cooperation. This, as we saw, is a crucial step in averting the "hazards of the generalized prisoner's dilemma" (*TJ*, p. 577/505). When we look at why Rawls thinks Joan's reasons to be just tell decisively in favor of doing this, we shall see that, in appealing to Joan's desire for ties of friendship, Rawls is using a commonly recognized strategy for avoiding prisoner's dilemmas. The weakness of the reason to which I am now drawing attention implies that that strategy has its limits, as Rawls acknowledges. That, as we shall see, is why he also needs congruence arguments that appeal to other considerations—specifically to the desire referred to by C_4a, the desire to express our nature as free and equal rational beings.

Perhaps the second argument will seem to establish stronger reasons if read in conjunction with the first argument. If, as the first argument supposes, Joan has an aversion to hypocrisy and deception, then she may be dissuaded from acting unjustly by the fact that passing on the gains of cheating is "dubious and involved." So perhaps Rawls means to draw on that aversion, together with Joan's desire to protect her friends, to show her reasons to

preserve her sense of justice are "strong." And if we read the second argument as taking for granted the aversion assumed in the first argument, then we can see why Rawls makes the undefended—and, in the text, the unstated—assumption that "natural and simple is best." This reading of the second argument also explains why Rawls begins the second argument by asserting a connection with the first.

I shall return to this suggestion in §V.5. Whether or not it is correct is beside the two points I hoped to make by going through the argument with care:

First, I hoped to show just exactly how this argument builds on groundwork laid much earlier in *TJ*—specifically, in *TJ*, §74, on moral and natural attitudes. Seeing this heightens appreciation for the unity of part III of *TJ* and enables us to see just how precisely that groundwork is laid. At *TJ*, pp. 485–86/425, Rawls says:

> in examining a moral feeling, we should ask: to what natural attitudes is it related? Now there are two questions here, one the converse of the other. The first asks about the natural attitudes that are shown to be absent when a person fails to have certain moral feelings. Whereas the second asks which natural attitudes are evidenced to be present when someone experiences a moral emotion.

He adds immediately that he has "been concerned only with the first question, since its converse raises other and more difficult problems." Perhaps the converse *does* raise more difficult problems. But surely another reason Rawls was concerned with the first question and not the second is that the first question was the one he needed to answer to establish a crucial premise in the argument I have just laid out—the argument from C_4c. That is the claim that "among persons who never acted in accordance with their duty of justice except as reasons of self-interest and expediency dictated,"—persons such as Joan thinks about becoming—"there would be no bonds of friendship or mutual trust" (*TJ*, p. 488/427).[9]

Second, I wanted to draw on the connection with earlier material in *TJ* to confirm my interpretation of congruence. I argued in Chapter II that the question of congruence is not a question about what goes on, as it were, act-by-act, nor is it a question about the characteristic motive of just acts. As if to confirm that point, and to anticipate the argument from C_4c, Rawls concludes *TJ*, §74 by remarking that:

> the fact that one who lacks a sense of justice, and thereby a liability to guilt, lacks certain fundamental attitudes and capacities is not to be taken as a reason for acting as justice dictates. But it has this significance:

9. As Rawls says of Locke's defense of a negative criterion of legitimacy: "very sensibly, he argues for what he needs and not more." See Rawls, *Lectures on the History of Political Philosophy*, p. 131.

> by understanding what it would be like not to have a sense of justice—that it would be to lack part of our humanity too—we are led to accept our having this sentiment. (*TJ*, p. 489/428)

The Desire to Participate in Forms of Life That Call Forth Talents

Of the four arguments Rawls offers in the passage with which I am now concerned, the third—the *Social Unions Argument*—may be the most difficult to make out. Certainly it is the most difficult for someone who wants to read into these arguments the strategy I have attributed to Rawls, for it seems hardly to square with that strategy at all. The argument does not seem to appeal either to the diversity of descriptions or to C_4d. But I think the *Social Unions Argument* can quite plausibly be read as relying on the strategy I have imputed to Rawls. In reading this argument, we need to bear in mind Rawls's remark that the chain of arguments in which this argument is a link "review[s] various points already made" (*TJ*, p. 570/499). The points being reviewed in this argument, found in §86 of *TJ*, are those already made in *TJ*, §79 on social unions.

I said in §III.4 that *TJ*, §79 falls into two parts. The preface, which I analyzed in some detail, treats of social unions generally without drawing on the principles of justice. It provides an argument for:

> C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

The second part of *TJ*, §79 then applies the points made about social unions generally to the special case of a social union of social unions, which is regulated by the principles of justice. On my reading of the argument about social unions in §86 of *TJ*, that argument largely presupposes the preface of *TJ*, §79 and recapitulates the points made in the second half of §79. I shall not go through the relevant part of §86 line by line, but I think the sequence of thought in the *Social Unions Argument* goes as follows.

Rawls assumes C_4d, which—as we saw—depends upon the Aristotelian Principle, together with the qualified version of the Companion Effect asserted at the step I referred to as (4.5'). The heart of the argument from C_4d is the passage in which Rawls says that a social union of social unions:

> realizes to a preeminent degree the various forms of human activity; and given the social nature of humankind, the fact that our potentialities and inclinations far surpass what can be expressed in any one life, we depend upon the cooperative endeavors of others not only for the means of well-being but to bring to fruition our latent powers. And with a certain success all around, each enjoys the greater richness and diversity of the collective activity. (*TJ*, p. 571/500)

Thus in a social union of social unions, the goods available in social unions generally are available "to a preeminent degree." This is because in a social union of social unions, our latent powers are brought more fully to fruition

than in smaller social unions, such as clubs and teams, and the diversity of activity is richer. From this, together with C_4d, we are supposed to infer that all members of the WOS have an especially strong interest in participating in a social union that includes all the smaller social unions.

In writing *TJ*, §86, I believe Rawls assumes we will recall an important claim from §79: the claim that to enjoy the goods of any social union, we must, as it were, play the game in the right spirit—we must participate in the technical sense of that term, valuing the activity for its own sake and affirming its common aim. The same, he argues here, is true of a social union of social unions. Since participation in a social union requires taking its rules and norms as regulative, participation in a social union of social unions requires taking the principles of justice as regulative. So if all members of the WOS want to participate in a social union of social unions, then Joan has another reason to preserve her sense of justice as a highest-order regulative desire in her rational plans.

This last bit of argument assumes that what is true of a social union is true of a social union of social unions. While this is true by definition, it is still illuminating to look more deeply. We know from the second half of §79 that the "shared final end" of a social union of social unions is "the successful carrying out of just institutions" (*TJ*, p. 527/462). So if the parallel between social unions and a social union of social unions holds, then enjoying the goods of a social union of social unions must depend upon members of the WOS affirming that aim by taking the principles as regulative.

And Rawls argues that it does. To see how *that* argument goes, it is useful to recall Rawls's game analogy. We saw that taking the rules of a game as regulative is itself experienced as a good, in accord with the second conjunct of the Aristotelian Principle. This is because taking them as regulative requires players to engage in the complex activity of devising modes of play that advance their interests consistent with the rules. These modes of play might include strategies that demand complex coordination with other players, feints, bluffs, creative reinterpretations of rules of play, or novel ways of executing familiar moves. Similarly, Rawls says, in a social union of social unions, "the plan of each person is given a more ample and rich structure than it would otherwise have; it is adjusted to the plans of others by mutually justifiable principles" (*TJ*, p. 528/463).

The relevant similarity between a game and a social union of social unions may be hard to detect, since in the latter, activities are heterogeneous in the extreme. The talk of strategies for winning within the rules may misleadingly suggest that the relation among social groups in the WOS is competitive. But some groups do compete for members and for public and private support. Moreover, each group has to adjust to the fact that its members belong to other groups which influence the spirit and regularity with which they take part. Here we need only think of religious organizations which, in the WOS, will be unable to insulate their members from the influences of groups with diverse membership and different ideas of how to live. This illustrates the need for the

mutual adjustment to which Rawls refers. However challenging this mutual adjustment may be, "this collective activity," Rawls thinks "if the Aristotelian Principle is sound, must be experienced as a good" (*TJ*, p. 528/463).

Another feature of the game example suggests a further line of argument. We saw that someone must participate in a game to appreciate the excellent play of others as realizations of his own nature and as something he has helped to call forth. So also he must participate in a social union of social unions to see the activities of others as a realization of his own nature and as something he has helped to elicit. Thus Rawls says "to appreciate something as ours, we must have a certain allegiance to it" (*TJ*, p. 571/500). The kind of allegiance he has in mind is "acknowledg[ment of] the principles of its regulative conception" (*TJ*, p. 571/500). But participation is not just necessary. For if someone participates, and does see the activities of a social union of social unions as in these ways his, then he may appreciate them, not by the Companion Effect, but—as I remarked in §IV.2—by the Aristotelian Principle itself. Hence in a social union the Aristotelian Principle itself has "its wider effect" (*TJ*, p. 571/500).

I think what Rawls has in mind is this. If members of the WOS do their part in upholding just institutions, and they know that just institutions make it possible for others to pursue their good—whether it be baseball or music, stamp collecting, or family life—then each person can see the pursuits of others as developments of her own latent abilities that *she* has helped to make possible. The person who likes music but devotes herself to sports can, for example, take some joy in the musical accomplishments of others because by developing their talents, others have developed their common human nature. She can also take some pride in those accomplishments because she knows she is part of a society that makes it possible for people to cultivate musical talent. Those musical accomplishments of others are, in both of these ways, her accomplishments as well. The two conjuncts of the Aristotelian Principle imply that she will experience those accomplishments as good.

Thus on my reading of the *Social Unions Argument*, the argument employs the strategy found in the first two arguments. Rawls assumes C_4d at the beginning of the argument. The value of members of the WOS attach to the end it refers to is accounted for by the thin theory. Rawls then argues that, because a social union of social unions "realizes…the various forms of human activity" "to a preeminent degree," the desire asserted by C_4d is best satisfied by participation in a social union of social unions. The argument therefore connects participation in a social union of social unions with reasons provided by the thin theory. The sense of justice is "connect[ed] with reasons specified by the thin theory of the good" by the nature of participation. Participation in a social union of social unions requires members of the WOS to "acknowledge the principles of its regulative conception" because of what participation is. And so it requires them to take the principles of justice as supremely regulative. When others take those principles as similarly regulative, Joan will therefore have as weighty a reason to preserve her desire to act from the principles of justice as she has to participate in social unions.

This connection between the sense of justice and reasons specified by the thin theory depends upon the fact that the principles satisfy a *diversity of descriptions*. They are principles of justice for regulating the basic structure of a WOS and they are the supremely regulative principles of a social union of social unions. Why say, as Rawls does, that "*the theory of justice* supplies other descriptions of what the sense of justice is a desire for" (*TJ*, p. 569/499, emphasis added) and that "*the details of the contract view*" (*TJ*, p. 571/500, emphasis added) establish the connection between the Aristotelian Principle and the goodness of participation in a social union of social unions? One reason is that the concepts of participation and of a social union are theoretical concepts. But another has to do with the fact that Rawls's theory is specifically a social contract theory. The principles regulate a social union of social unions because they are chosen in a contract subject to the condition of finality imposed in *TJ*, §23. That condition requires that the principles, like the rules of a game, serve as the final court of appeal for settling conflicts.

The Desire to Express Our Nature

The fourth argument, which is the most philosophically interesting, is the most seemingly straightforward. In fact I think the argument is quite complicated, and I shall defer any sustained analysis of it until Chapter VII. Here I shall simply present the bare bones of the argument as Rawls lays it out in *TJ*, §86. Because I shall subsequently want to draw attention to various moves in the argument, I shall lay it out in premise-and-conclusion form.

Rawls assumes that all members of the WOS have the desire asserted by C_4a, which says:

> C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

He asserts that:

> (5.2) The desire to express our nature is a desire to act from principles that would be chosen in the OP.

Clearly

> (5.3) The desire to act justly is the desire to act from the principles that would be chosen in the OP.

(5.2) and (5.3) imply that:

> (5.4) The desire to express our nature has the same object as the desire to act justly.

It follows that:

> (5.5) Joan can satisfy the desire asserted in C_4a by and only by acting justly.

It seems that Joan has at least as much reason to act justly as she does to satisfy the desire to express her nature.

We have seen that the conclusions Rawls ultimately wants to establish concern judgments Joan would make about her balance of reasons. To make such a judgment, it is not enough that (5.5) be true, that Joan has reasons to act justly, or that those reasons are at least as strong as her reasons to express her nature. Joan must see that she has such reasons and she must see how strong they are. And so she must know that (5.5) is true. But of course, she does know that. For Rawls says that "we are concerned only with the special case of the well-ordered society as characterized by the theory." In that case, the publicity condition is satisfied and justice as fairness is public knowledge. Moral education is transparent, so that everyone "come[s] to know the derivation of moral precepts and ideals" (*TJ*, p. 496/434; cf. *TJ*, p. 515/452). So:

(5.6) "we are entitled to assume that [the] members [of the WOS] have a lucid grasp of the public conception of justice upon which their relations are founded." (*TJ*, p. 572/501)

And this implies that Joan, like everyone in the WOS, knows (5.5). Since she knows that her desire to express her nature can only be satisfied by acting from principles of justice:

(5.7) Joan's desire to express her nature moves her to act justly.

This is why Rawls says:

(5.8) "when someone has true beliefs and a correct understanding of the theory of justice, these two desires move him in the same way." (*TJ*, p. 572/501)

Thus for members of the WOS, and hence for Joan, the desires to express one's nature and to act justly are identical in practice. As Rawls puts it:

(5.9) "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*, p. 572/501)

The desire to act justly and the desire to express our nature "are both dispositions to act from precisely the same principles: namely, those that would be chosen in the original position" (*TJ*, p. 572/501). It follows that Joan has just as much reason to act from those principles as she has to express her nature.

In this argument, much of the strategy I have imputed to Rawls is readily apparent. He assumes C_4a at the beginning of the argument. He moves from C_4a to (5.5) by drawing on contract theory to say what the desire to act justly is a desire for, the description being asserted in (5.3). What Rawls calls the "practical identity" asserted in (5.4) and (5.5), and reiterated in (5.9) connects the sense of justice with reasons of Joan's that are specified by the thin theory at C_4a (*TJ*, p. 572/501).

§V.4: Some Questions about the First Three Arguments

I have tried to show that the desires asserted by C_4a, C_4b, C_4c, and C_4d are best satisfied when members of the WOS affirm that being just is part of their good, and resolve to maintain their sense of justice. These arguments show reasons that Joan has to be just. They do not depend upon Joan's having ideal-dependent desires to live up to the ideals of justice as fairness, or upon her having any other desires for objects the value of which is given by the full theory of the good. Rather, they are supposed to show the reasons that Joan has to maintain her sense of justice insofar as she "follows the thin theory." Identifying these reasons is the first stage of the two-stage argument for congruence.

In the second stage, Rawls argues that these reasons are decisive. I shall begin looking at those arguments in Chapter VI. But as I have suggested, even in laying out the reasons, Rawls conveys the clear impression that he thinks the reasons are strong ones. I now want to look at what Rawls thinks the strength of those reasons depends on. I shall argue that the reasons identified by the first three arguments draw their strength from a common source. Locating that source deepens appreciation for the central role of the Aristotelian Principles in Rawls's treatment of congruence. It also brings to light one of the features of the congruence arguments with which Rawls became dissatisfied. So let me now turn to Rawls's implication that Joan has strong reasons to preserve her sense of justice.

Rawls is surely right to maintain, in the first argument, that Joan can be sure of *never* having to pay the costs of hypocrisy only if she treats her sense of justice as supremely regulative. But Joan will take this to be a strong reason only if she is strongly averse to paying those costs. The question is why she would be. There are a couple of possibilities. She would be strongly averse if she regarded the costs as intolerably high in absolute terms. She would also be strongly averse if she regarded the costs as high relative to what she thinks she could get by paying them. The former seems unlikely. So I suspect Rawls's implication that the first argument identifies a strong reason depends upon the latter claim. It depends, that is, on Joan's treating the costs of hypocrisy as high relative to the benefit of, for example, the greater wealth she might enjoy by cheating on her taxes.

But if Rawls's first argument does depend upon this, then it requires a further argument that Joan will not attach especially high value to wealth above her fair share. If that further, supplemental argument is to be of use, it cannot allege that Joan would not value wealth above her fair share because she is troubled by the prospect of acting unfairly, since we are supposing that Joan follows the thin theory and are asking why she should affirm her sense of right. Rather, the argument must be that she does not care that much about the extra wealth at all. In *TJ*, §82 Rawls offers an argument that seems to provide the supplement needed by the argument now under consideration—the argument premised on C_4b. The argument in §82 purports to show that one

reason members of the WOS might be thought to have for seeking extra wealth, namely status-seeking, would not in fact move them.

Let me just note two points about that argument.

One is that, if the argument of §82 is indeed needed to supplement the argument premised on C_4b, then that is further evidence that Rawls's treatment of congruence draws together considerations from elsewhere in chapter 9, though in this case the dependence is not signaled by any obvious cross-references in the text.

Second, the argument of *TJ*, §82 depends upon Rawls's assertion that in the WOS, "the position of equal citizenship answers to the need for status." Rawls's idea seems to be that the desire for status that sometimes manifests itself in a desire for wealth is in fact a desire for the grounds of self-respect. He suggests as much in his essay "Fairness to Goodness" where he says "strong or inordinate desires for primary goods on the part of individuals and groups, particularly a desire for greater income and wealth and prerogatives of position, spring from insecurity and anxiety."[10] His reply in *TJ* seems to be that if Joan knows that others respect her as an equal citizen, then she will have the grounds of self-respect that she needs. She will not, therefore, be moved to seek wealth, possessions, or relatively high economic status as a means to self-respect. Thus, the fundamental assertion in Rawls's argument is, we might say, that "the position of equal citizenship answers to the need *we might have thought people had* for *economic* status." If this assertion is right, then it must be that Joan will not be troubled if others have more than she, at least if inequalities are not excessive. But why won't she be? Why, exactly, does Joan's position of equal citizenship "answer[ ] to [her] need for status"? Why should Joan attach that kind of value to her standing as an equal citizen?[11]

One possibility is this: When Joan observes that someone has greater wealth than she, she thinks to herself something like "No matter. I am his equal in the way that counts because the extra things are just so many empty trifles compared to my liberties and opportunities." The idea is that equality of liberty and opportunity is what really matters to people because liberty and opportunity are much more valuable than wealth. When they are and are known to be equal in what is most valuable, no one feels any need to look elsewhere for self-esteem. And since equal citizenship is what confers equality in what really matters, equal citizenship is what answers to the need for status.

It may be appealing to impute this explanation to Rawls, since in arguing for the priority of the first principle to the second, he appeals to the claim that does the explanatory work: the claim that it is rational to value liberties over

10. Rawls, Collected *Papers*, p. 277.

11. Joshua Cohen, "Taking People as They Are," *Philosophy and Public Affairs* 30 (2001): pp. 363–86 includes a helpful presentation of the argument at p. 382. Cohen's article shows just how important this argument is to Rawls's ability to fend off a powerful objection. Because Cohen's interests lie elsewhere, he does not pursue the question I raise here.

income and wealth. Even so, it is unlikely that this is the explanation Rawls has in mind. For people who are equal in other respects sometimes exalt differences that it would be rational to regard as trivial in order to distinguish themselves. We need to know why that does not happen in the WOS. More specifically, we need to know why members of the WOS do not look to differences in wealth as grounds for the sort of social distinctions that undermine self-respect, even if it would be rational for them to regard differences in wealth as inconsequential.

While I cannot provide a full answer here, I do want to zero in on one strand in the answer Rawls would give that may be overlooked. In the WOS, where the principles of justice are satisfied, where good education and training are available, where talents are widely dispersed, and where intergenerational transfers are limited, differences in wealth will in large part be the result of members' exercise of choice. When members of the WOS see others who are better off than they, they can tell themselves that they had ample opportunities to pursue more lucrative occupations, but chose not to do so because they thought they would find satisfaction in the plans they in fact adopted. Differences in income and wealth are, Rawls may think, unlikely to undermine the self-respect of the less well-off when everyone recognizes that the less well-off could have had more. And in the WOS, everyone recognizes that the less well-off could have had more because citizenship—hence liberty and opportunity—are equal.

This is only a partial answer, since differences in talent may lead to some disparities of wealth in the WOS. But if the partial answer is right, then the connection between one's choices and one's social position goes some way to explaining why Rawls thinks that "equal citizenship answers to the need for status," at least among the equally talented. It therefore goes some way to explaining why Joan does not care all that much about extra wealth. And it therefore goes some way to explaining Rawls's otherwise puzzling assumption that Joan would regard the costs of hypocrisy and deception as high relative to the benefits of the wealth she could get above her fair share. So the strength of the first reason Rawls identifies for Joan to treat her sense of justice as supremely regulative—the reason identified by the argument from C_4b—seems ultimately to depend upon Joan's finding satisfaction in knowing that she is living the life she has chosen.

Why should Joan find the fact that her life is chosen satisfying enough to do this work? The answer, I think, is that she finds satisfaction in that fact because C_4a is true of her—because she thinks of herself as a free rational person, and wants to frame and live out her life as such a person. Thus, the strength of the reason identified by the first argument ultimately depends upon people's thinking of themselves as free, and upon their desire to live freely being such that when the desire is satisfied, they do not want other things badly enough to act against the view they have of themselves. It therefore depends upon members of the WOS having the *free-and-equal self-conception*, and the desire to be live up to it.

Now consider the second argument. Rawls is surely right to say that Joan can be sure of protecting her friends from the consequences of her own injustice

only if she successfully resolves to preserve her sense of justice. But here, too, it may seem that the reasons to treat justice as regulative are weaker than Rawls allows. Imagine that Joan considers the possibility of passing along more wealth to her children than they should receive. She knows that doing so may require some deception and hypocrisy on her part, and that it runs the risk of hurting other persons and institutions she cares for. She has some interest in protecting those persons and institutions. But Rawls's claim is that the potential costs to those with whom she has somewhat distant relations give her strong reasons to treat her sense of justice as supremely regulative. This conclusion surely depends upon his assumption that she will judge these costs high relative to the benefit she can confer on her children. Why should she do that?

I think Rawls would respond that what Joan and other members of the WOS want for their children is that they be able to choose and live out lives they find satisfying. Joan and the other members of the WOS assume that, as with them so with their children, citizenship goes some way to answering the need for status. They know that their children no more need wealth for self-respect than they do themselves. In a society with fair equality of opportunity, even access to high-quality education commensurate with extraordinary talent will not require greater resources than justice allows. Members of the WOS know that their children, like all members of the WOS, have the liberties, opportunities, and resources they need to choose and live satisfying lives. They therefore care relatively little about being able to pass on more—where "care relatively little" entails caring less about passing along more than about the costs of doing so. If this is Rawls's answer, then we can see why he thinks the reason identified by the second argument—the argument premised on C_4c—is strong. The strength of the reason depends upon Joan's thinking that her children do not care about economic status any more than she does, and that they have the kind of lives she values for them because they have the lives they choose. It therefore depends upon Joan's attaching very high value, now not to the fact that she herself lives as a free rational agent, but to the fact that her children live as such. That, she thinks, is their nature as it is hers. And as the expression of her own nature is satisfying to her, so she thinks, their expression of their nature will be satisfying to them.

Consider, finally, the reason identified by the *Social Unions Argument*. That reason, too, is a reason for Joan to maintain her sense of justice as supremely regulative. It is thus a reason for her not to treat it as a sentiment she can act against, even if—for example—she is tempted to try restricting the liberties of groups whose activities she finds offensive. She may think the offense she takes at a group's religious or sexual practices provides her some reason to try to repress them, but Rawls seems to think the reason she has to take her sense of justice as supremely regulative is stronger. It *would* be if Joan regarded the cost of not participating fully in a social union of social unions as a high cost relative to what might be gained by repressing an offensive group. But why would she? Why would she attach relatively little value to the repression of offensive activity?

Rawls would answer, I suspect, that Joan she does not just value participation in a social union because she "enjoys the greater richness and diversity of collective activity" and finds it satisfying to take part in eliciting that diversity. She also values taking part in a collective life that makes it possible for others to live as persons who are free in the sense that they are living the lives they choose. This suggests that the argument for C_4d is more complicated than it first appeared. As I reconstructed that argument in §IV.2, it depended upon the qualified version of the Aristotelian Principle's Companion Effect:

> (4.5') "When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept" (*TJ*, p. 523/459).

If the reply I am now exploring on Rawls's behalf is right, then the disposition to enjoy the perfections of others does not just depend upon their developing a variety of talents. It also depends crucially upon their choosing which ways of life to pursue. The upshot is that even if Joan finds the lives they choose as in some way offensive, she knows that she is taking part in a form of life that lets them express their nature as choosers. If she takes sufficient satisfaction in that, then she *will* have a strong reason to affirm her sense of justice.

The presence of the desires asserted in C_4c and C_4d—hence the arguments that depend upon them—clearly depend upon the Aristotelian Principle and some form of the Companion Effect. Moreover, we saw in §IV.2 that the presence of a desire to express our nature is explained by the Aristotelian Principle and, in particular, by its first conjunct, which grounds the claim that we have a desire to exercise our natural faculties. So the fourth argument—which is premised on C_4a, the claim that we have desire to express our nature—also depends upon the Aristotelian Principle. If I am right about why Rawls thinks the reasons identified by the first three arguments are strong, then all four arguments depend upon members of the WOS thinking of themselves and one another as having the nature (1.1) and the *free-and-equal self-conception* imply they do: a free nature. And all four arguments depend upon their attaching great value to the expression of human nature, either in their own lives or in the lives of others. We shall see in Chapter VIII that different senses of "freedom" are at work in the arguments. For now, note that the Aristotelian Principle explains the presence of the desire to express our nature; the Companion Effect says that we take pleasure in the excellences of others. If living freely is itself an excellence, then the Principle and its Companion Effect support all four of the reasons to be just that Rawls identifies.

Seeing how these reasons depend upon the Principle and the Companion Effect, we can now see why Rawls implies that his treatment of congruence depends upon the Aristotelian Principle in the passage that I said Freeman found so suggestive in §V.2, the passage in which Rawls says the claim that living in a just society is a good depends upon the Aristotelian Principle. Seeing just how the Principle and the Companion Effect are drawn on in the

original arguments for congruence also sets up a useful comparison with the way that Rawls later discusses the good of a WOS. I shall argue that seeing what implications of the Principle and the Companion Effect are—and are *not*—drawn on in the later treatment confirms the interpretation I shall offer of Rawls's later dissatisfaction with his original treatment of congruence.

§V.5: Some Puzzles about the Fourth Argument

Rawls's implication that the reason identified by the fourth argument is decisive is puzzling for different reasons.

First, unlike arguments from C_4b, C_4c, and C_4d, the argument from C_4a is not conditional on the assumption that others are just. Rather, it seems to show that Joan has a reason to act from the principles of justice regardless of how others behave. This conclusion in itself is not problematic. In fact, the conclusion is intuitively plausible. But it would take a very powerful argument to show that the reason that depends on C_4a is decisive, since it would take a very powerful argument to show that we have *decisive* reason to act from the principles of justice regardless of how others treat us. Moreover, such an argument would seem to show more than Rawls needs to show in order to meet the challenge his treatment of congruence is supposed to address. For as I have stressed, Rawls thinks the real challenge is that of showing that a just plan of life is Joan's "best reply" in the special case in which others make "similar plans" (*TJ*, p. 568/497). It seems, then, that the argument from C_4a picks out what is, for Rawls's purposes, the wrong reason.

The problem seems to be due to (5.9) and the steps that support it. For example, (5.7) says:

(5.7) Joan's desire to express her nature moves her to act justly.

Perhaps this step is too strong, since all Rawls seems to need is weaker claim that results from adding the italicized reciprocity rider: "Joan's desire to express her nature moves her to act justly *when others act justly as well*." Although an argument that goes by way of this weaker claim might be enough, we shall see in Chapter VII that Rawls does not weaken the premises of his argument this way.

But if (5.7) seems too strong because it does not include the reciprocity rider, there is another respect in which it seems too weak. The arguments for congruence are ultimately supposed to show that members of the WOS would live a certain sort of life: the life of a person who not only acts justly, but also of a person whose higher-order desire to be a just person is not outweighed by competing desires, however strong. To show this stronger conclusion, what Rawls really needs is a claim about, not about how Joan treats the principles, but about how she treats her desire to act on the principles—a claim, that is, about her sense´ of justice. More specifically, what Rawls needs is not (5.7) but:

(5.7') Joan's desire to express her nature moves her *to treat her sense of justice as supremely regulative of her other desires.*

If Rawls could defend (5.7'), he could move – via (5.8)—to:

(5.9') "The desire to [treat our sense of justice as supremely regulative] and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire."

The crucial claim in the argument for (5.7) is

(5.5) Joan can satisfy the desire asserted in C_4a by and only by acting from her sense of justice.

Rawls could get to (5.7') instead of the weaker (5.7) if, instead of (5.5), he could show that:

(5.5') Joan can satisfy the desire asserted in C_4a by and only by treating her sense of justice as supremely regulative of her other desires.

In Chapter VII, we will see how Rawls gets to (5.5'), and hence to (5.7') and (5.9'). He can then infer that Joan has as strong a reason to treat her sense of justice as supremely regulative as she does to express her nature. Even this may not seem to be as strong a conclusion as Rawls needs, since what Rawls wants to show is that the desire to express her nature gives Joan reason to *preserve* her sense of justice as supremely regulative. In Chapter VII, we will also see how Rawls takes this last step.

There is one last puzzle about the argument from the desire to express our nature to which I want to draw attention. The argument depends upon (5.2), the claim that our desire to express our nature is a desire to act from principles that would be chosen in the OP. But it is by no means evident that this is so. In the course of saying that Rawls does indeed accept (5.5'), I will also show why he accepts (5.2). Showing all this will ultimately cast further light on the connections between §86 of *TJ* and its immediate context, the sections on hedonism, dominant ends, and the unity of the self. It will therefore enable us to see why—as I promised in §II.3—Rawls thinks everyone's affirming her sense of justice stabilizes the WOS by giving unity to practical reason. Seeing this, in turn, fills in more details of the conception of themselves and others held by members of the WOS. For it shows part of what is involved in their thinking of themselves and others, not just as free, but as practically rational.

VI

The *Argument from Love and Justice*

I have said that Rawls made the changes he did between *TJ* and *PL* because he became dissatisfied with *TJ*'s treatment of congruence. In support of this interpretation, I have shown what Rawls means by congruence and why he thinks he needs to show it. In Chapter V, I reintroduced the case that poses the fundamental problem of congruence—namely, the case of Joan—and went through the four reasons that Rawls thinks Joan has to maintain her desire to act from the principles of justice. I detailed the strategy Rawls relies upon to show that Joan has those reasons. He starts with desires for ends the values of which are given by the thin theory, and argues that those desires can best or only be satisfied if Joan maintains her sense of justice. Joan therefore has thin reasons to maintain that sentiment.

I raised some questions about the strength of the first three reasons, and asked whether the fourth is really a reason for Joan to do what Rawls thinks she has reason to do. Despite the questions that can be raised about those reasons, the Rawls of *TJ* insists that they tell decisively in favor of congruence. Joan's balance of thin reasons, he thinks, tilts in favor of maintaining her desire to act from the principles of justice when others maintain theirs. If Joan also knows that everyone else also has a sense of justice and that each person's balance tilts in the same way hers does, she will affirm her sense own sense of justice from within both the thin theory and the viewpoint of full deliberative rationality. This solves the congruence problem in its non-trivial form.

In this chapter and the next, I want to look closely at Rawls's arguments for these conclusions. Here as in Chapter V, I shall follow the order of his text. The treatment of what I call the *Argument from Love and Justice*, which begins at *TJ*, p. 573/502, is therefore deferred until §VI.4. It is preceded by substantial expositions of Rawls's own prefatory—and cursory—remarks on methodology, remarks that apply both to the congruence argument I shall look at in this chapter and to the one I shall examine in the next. These methodological remarks are very abstract, but the later treatment of the *Argument from Love and Justice* will, I hope, eventually make them less so.

There are a number of reasons for attending to the details of Rawls's congruence arguments. For one thing, the arguments are easily misread, and commentators sometimes mistake exactly what they are supposed to show and exactly what the relationship is between them. Even close readers then miss what Rawls later came to find unsatisfactory about them. Furthermore, while Rawls is sometimes read as having dismissed intuitionism early in *TJ* to focus on utilitarianism, seeing how the congruence arguments go shows that Rawls is profitably read as fighting a two-front war against both of these philosophical views. Correct interpretation of the *Argument from Love and Justice* also shows that criticism of justice as fairness as an individualistic doctrine badly caricatures Rawls's view. Finally, as we shall see in Chapter VII, a correct reading of the *Kantian Congruence Argument* deepens our appreciation of Rawls's debt to Kant. It also shows how Rawls uses the OP to "bridge" the right and the good. The fact that the OP plays what I call the *bridge function* bears on the question of whether the OP is an intellectual device that is, in principle, dispensable.

§VI.1: Balances and Temptations

After arguing that the typical member of the WOS Joan has the four reasons to be just that I surveyed in Chapter V, Rawls says

> Let us suppose that these are the chief reasons (or typical thereof) which the thin account of the good allows for maintaining one's sense of justice. The question now arises whether they are decisive. Here we confront the familiar difficulty of the balance of motives[.] (*TJ*, p. 572/501)

I take the last sentence of this brief passage to indicate that Rawls is going to show the reasons "which the thin theory of the good allows" are decisive by showing that Joan would judge, from within the thin theory, that her balance of reasons tilts in favor of "maintaining [her] sense of justice."

That stability should depend on each person's balance of reasons is only to be expected. If we have a sense of justice, our practical reasoning must take account of its demands as well as of our other desires. Some of our desires will move us in the same direction as our sense of justice does. But the task of deciding what to do, and how to plan our lives, is complicated by the fact that we face temptations to act unjustly. There is no such thing as a person who

does not face temptation. So whether we are just or not depends upon how we cope with these competing desires. Rawls indicates as much when he says:

> 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. To estimate the stability of a conception of justice (and the well-ordered society that it defines), one must examine the relative strength of these opposing tendencies. (*TJ*, p. 454–55/398)

This passage from early in *TJ*'s treatment of stability lends further support to my claim that Rawls's congruence arguments concern balances of reasons, and against interpretations that depend upon reading his conclusion differently.[1]

We shall begin to see in §VI.3 that my reading also fits better with how the arguments actually go than do alternative interpretations. As I suggested in §V.2, I believe that subtle differences in how the conclusion of the congruence arguments is to be worded reflect deep and important differences in how the point of those arguments is to be understood. I read the congruence arguments as attempts to establish a conclusion about the balance of reasons because I read them as attempts to show something about how the rewards of various courses of action compare. I read the arguments this way because I take them to respond to a specific threat to the stability of justice as fairness: in light of her "propensities to injustice," each member of the WOS might think it rational to act against her desire to act from the principles of justice and defect from the agreement that would be reached in the OP. Readings according to which the conclusions of the congruence arguments do *not* concern the balance of reasons typically overlook the game-theoretic threat to which the arguments respond and the value the just person attaches to her other ends. By doing so, they overlook the great ambitions of Rawls's attempt to show the inherent stability of justice as fairness—ambitions I discussed in §§II.3 and III.4.

But if temptations to injustice are a fact of Joan's life, it is important not to mistake the way this fact enters into her decision about whether to affirm her sense of justice. Let me therefore review the choice Joan faces.

The sense of justice is a trait of character that Joan is assumed to have because she has grown up in a WOS. The question of congruence does not concern two acts she might perform at any given time, one dictated by her sense of justice and the other unjust, such as the question of whether to pay her taxes or to cheat. Rather, her alternatives are two different kinds of person she might be or two different lives she might live. It is important that those two lives are not the lives of the just person and the habitually unjust one. One *is* the life of the just person, the person who takes her desire to act from the principles of justice as supremely regulative and who tries to preserve that sentiment. But the other is the life of the person who knows what justice

1. See Barry, "Search for Stability," p. 887.

requires, but who decides case-by-case whether to act justly, even when others treat their desire to act from the principles as supremely regulative. So one is the life of a person who is just, come what may. The other is the life of a person who takes considerations of justice into account, but who decides what to do by weighing them against other desires she has. The decision Joan faces is a fundamental decision about her plan of life, about the kind of character she will have, and the kind of person she wants to be. Thus on my reading, *TJ*'s discussion of life-plans lays the groundwork for Rawls's explicit treatment of congruence, even though it is to be found almost 150 pages earlier.

The question Joan asks herself about what sort of person to be is one that she can pose at any point in her life, for at any time, she can ask whether she is glad she has her sense of justice and is glad she treats it as she does. As we shall see, in trying to answer the question, she will have to consider whether she will later have cause to regret her choice. Just as the discussion of plans of life earlier in *TJ*'s anticipates the statement of Joan's alternatives, so too that discussion anticipates the solution to her choice problem, with its attendant questions about balances of reasons and future regrets. For Rawls says "a rational individual is always to act so that he need never blame himself no matter how his plans finally work out. Viewing himself as one continuing being over time, he can say that at each moment of his life he has done what the balance of reasons required, or at least permitted" (*TJ*, pp. 422/370–71).

Thus among the things Joan has to consider is whether, at some future time, she will look back and ask if—at the times she decided to preserve her sense of justice—she did "what her balance of reasons required, or at least permitted." Joan is a typical member of the WOS. The questions she asks herself are questions anyone can ask. The answers she arrives at are answers everyone will reach. If justice as fairness is to be stable, then whenever each member of the WOS asks about what sort of person to be, she must judge that her balance of reasons tips, and always has tipped, toward being just. This confirms that Rawls really wants to reach what I have called the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

I have also said that Rawls gets to that answer by way of *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

We have seen that C_N can be illustrated by Table II.3, where $A > B > D > C$.

Table II.3

	Player 2	
	Maintain regulative desire to act from the principles	Decide case-by-case
Maintain regulative desire to act from the principles	A, A	C, B
Player 1		
Decide case-by-case	B, C	D, D

This table resembles a prisoners' dilemma table, and the similarity raises an important question about Table II.3. In prisoners' dilemmas, the prisoners are assumed to care only about the length of their sentences. The values of various outcomes in prisoner's dilemma tables like Table II.1 are therefore expressed in the "currency" of jail time served. When I introduced Table II.3, I did not say much about how the values of the outcomes A, B, C, and D are to be expressed. The discussion of "thin reasons" in Chapter V enables me to answer this question.

We saw in §III.4 that if the congruence problem is to be solved in its non-trivial form, Joan must judge that her balance of reasons tips toward maintaining her sense of justice *from within the thin theory*. This, as we saw in §V.1, is why Rawls says that "the real problem of congruence is what happens if we imagine someone to give weight to [her] 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*, p. 569/499). This important remark about the weight Joan gives to maintaining her sense of justice implies that A and C, the values Joan attaches to being just, do not depend upon the fact that Joan wants to be just as such or upon her desire for other things the value of which is given by the full theory of the good. Rather, the values of A and C must depend entirely upon the weight she attaches to what I called her "thin reasons" to be just. More specifically, the values of A and C must be functions of just the values Joan attaches to other ends she wants and can get only by living as a just person, ends the value of which is accounted for by the thin theory of the good. Recall that in Chapter V, we saw what thin reasons Joan has to be just. They are the reasons she has because she has the desires referred to by:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

So to show that Joan's balance of reasons tips in favor of being a just person when others are just is to show is that A outweighs, or exceeds the value of B, *when A, like C, is reckoned just in terms of the value Joan attaches to satisfying these desires.*

My use of a payoff table to show how Rawls establishes C_N raises a number of philosophical and textual questions. I take up some of them in the next two sections, but those taken up in §VI.2 are, though important, rather arcane. Readers who are uninterested in the technicalities may want to skip directly to §VI.3.

§VI.2: Two Questions about Table II.3

Table II.3 and the discussion preceding it may seem to oversimplify the problem of congruence, for they suggest that when Joan and other members of the WOS ask themselves what sort of person to be, they have only two options or strategies. The table does not show what happens if Joan opts for a life of "mixed" strategies, sometimes treating her desire to act from the principles of justice as supremely regulative and sometimes, or in some kinds of cases, leaving herself the option of deciding how to behave. But there might seem to be many possible ways for Joan to combine the two strategies—many possible recipes for "mixing her life," as it were—and so Table II.3 might seem to require many more entries than the ones I have shown. This would pose a serious difficulty. Technically, a player's mixed strategy is a probability distribution over the pure strategies open to her—in this case, the two strategies depicted in Table II.3—reflecting the likelihood that she will play one pure strategy or the other.[2] So in making her decision about what kind of life to lead, each member of the WOS would have to ask how likely it is that others will mix their lives in all the various possible ways, and would have to consider which mixture would be her own best reply, all given that others are making the same calculations about her.

The possibility that members of the WOS could "play" mixed strategies would complicate the congruence problem enormously, and so make it very difficult to show inherent stability. Rawls makes some remarks that can easily be overlooked but that are, I believe, intended to rule out precisely this possibility.

By adopting the viewpoint of the thin theory, Joan adopts what may seem to be a somewhat artificial perspective on her desires. But despite the artificiality of her perspective, Joan is not an artificial person like the parties in the OP, whose psychology is open to stipulation. She is a real member of the WOS, typical in her desires and possessed of a sense of justice. Her behavior is governed by psychological laws and regularities. The sense of justice, as developed in the WOS, is a disposition to treat everyone justly, come what may. One of the regularities that governs Joan's psychology is this: the sense of justice, if affirmed, is an enduring trait of character "that…can be changed only gradually" (*TJ*, p. 568/498). So if Joan were to adopt the first strategy, and opt to maintain her sense of justice, she would thereby become the kind of person who could not

2. Robert Gibbons, *Game Theory for Applied Economists* (Princeton, NJ: Princeton University Press, 1992), p. 31.

then set aside this sentiment easily. Rawls says "we cannot preserve our sense of justice and all that this implies while at the same time holding ourselves ready to act unjustly should doing so promise some personal advantage" (*TJ*, p. 569/498). That is why, he says, the "just person is *not prepared* to do certain things" (*TJ*, p. 569/498, emphasis added). Mixing strategies—by treating the sense of justice as supremely regulative sometimes and sometimes deciding case-by-case—would require us to "hold[] ourselves ready to act unjustly." But if we do this, then we have not either affirmed our sense of justice or played a mixed strategy after all. Rather, in virtue of "hold[ing] ourselves ready to act unjustly," *we would have played the second strategy*. So in fact Joan has only two choices open to her, just as Table II.3 shows. This is what Rawls by saying in the same passage that "we"—and hence Joan—"cannot have things both ways" (*TJ*, p. 569/498).

While the foregoing argument may show that Joan cannot engage in random or arbitrary mixing, it does not show that she cannot engage in what we might call "principled mixing." Joan might consider being the kind of person who draws a principled distinction between those toward whom she will behave as a just person come what may, and those toward whom she will be more calculating. For example, she might consider being the sort of person who is just to her friends, her family, and her ethnic community and its associations, but who feels no such obligation when dealing with outsiders. Being this kind of person would require her to repudiate the form of her sense of justice that the WOS encourages, and so to reject the first strategy. But unlike arbitrary mixing, it does not require her to play the second strategy, or to be the second kind of person either. If principled mixing is an option for Joan, Table II.3 does not have enough entries.

To see how Rawls would respond, let us examine this "principled mixing" more closely. The kind of person Joan is now thinking about being is the kind of person who acts on a principle or maxim requiring her to maintain her sense of justice toward those persons and institutions she cares about. That is what distinguishes her from the arbitrary mixer. What distinguishes her from the person who affirms her sense of justice as encouraged by the WOS is that her circle of care is sharply limited. In Chapter V, we saw that in the special circumstances of the WOS, someone who thinks she has reasons to act on the maxim Joan is considering will have reason to maintain her sense of justice toward everyone and toward basic institutions. She will have those reasons because in the WOS, ties of friendship would extend so widely. And so if reasons to act on the principle or maxim are decisive, as Joan thinks they might be, then they tell in favor of maintaining her sense of justice rather than against it. I shall ask later just how persuasive this argument is. What matters for present purposes is that because Rawls thinks ties would extend so widely in the WOS, the strategy of "principled mixing" can be eliminated and need not be given a separate entry in Table II.3.

Does Rawls really try to show that members of the WOS face Table II.3 in its entirety? Or does my claim that he does read too much game theory into Rawls's discussion of congruence?

As I have said, Rawls does try to show that "the plan of life which [treats the sense of justice as supremely regulative] is [Joan's] best reply to the similar plans of [her] associates" (*TJ*, p. 568/497). This confirms that Rawls wants to compare the two entries in the left column of Table II.3. But Rawls never takes up the questions to which the right column provides answers, the question of how C and D compare and hence of how each member of the WOS should respond if everyone else decides not to affirm her sense of justice. Indeed, this is a question Rawls seems to dismiss when he begins laying out the reasons to be just that I discussed in Chapter V, for he says "others are assumed to have (and to continue to have) an effective sense of justice" (*TJ*, p. 570/499). This assumption seems to imply that the right column is irrelevant to the treatment of congruence. If it does, then my claims that Rawls tries to show members of the WOS face Table II.3, and that A > B > D > C, are mistaken or exaggerate the use to which elementary game theory illuminates the problem of congruence.

To address this worry, I want to look more closely into the assumption that "others are assumed to have (and to continue to have) an effective sense of justice."

On one reading of this assumption, when Joan asks herself what sort of person to be, she assumes that all other members of the WOS have already confirmed their sense of justice irrevocably, and she asks herself how best to respond to their irrevocable commitments. On this interpretation, Rawls says nothing about the right column of Table II.3 because the question it answers cannot arise.

But this reading of the assumption is mistaken. Joan is a typical member of the WOS. She is not asking herself a question that others are incapable of asking, and she knows that. So it is hard to see what grounds she could have for thinking that others have settled that question once and for all. Moreover, it is each person's knowledge that others can ask themselves that question that gives rise to what I called the *mutual assurance problem*—the problem of what assurance each member of the WOS can have that others will affirm their desire to be just. That is a problem that Rawls thinks his treatment of congruence has to solve, as I argued in §II.1. And, as I implied then, it is a problem he thinks is solved in part by showing how the payoffs of various courses of action compare. If Joan could assume that others had irrevocably committed to their sense of justice when she asks what kind of person she has most reason to be, the *mutual assurance problem* would already have been solved before the question about payoffs comes up.

The assumption that "others…have (and…continue to have) an effective sense of justice," and Rawls's silence about the relative values of C and D, must therefore be understood differently. On my reading, Rawls makes the assumption so that he can zero in on the left column. He zeroes in on that column, and ignores the right one, because comparing entries in the left column is all he thinks he needs to do. But the claim that this is all he needs to do itself rests on interesting claims about C and D, the payoffs in the right column. Rawls does not spell out those claims, and so I need to fill in what I take his reasoning to be.

I have said Rawls wants to show that members of the WOS have the strongest preference for mutual cooperation. Suppose that D, the payoff for being the kind of person who decides case-by-case, exceeds C, the payoff for affirming one's desire to act from the principles of justice even when others do not. On this assumption, mutual cooperation is preferred to the other possible outcomes if (i) A, the payoff for affirming one's sense of justice when others do, exceeds B, the payoff for replying to others' decision to be just by being the kind of person who decides case-by-case, and (ii) B exceeds D. The second of these conditions, (ii), seems obviously true. For it seems obvious that Joan would value B over D, since if others are just, she can take advantage of them if she decides case-by-case, while she cannot if they are just like she is. So on the assumption that D exceeds C, the interesting comparison—the one that needs to be drawn if Rawls is to show that mutual cooperation is the preferred outcome—is the one on which I have said Rawls focuses, the comparison of A and B.

But is it safe to assume, as I did for purposes of the argument I just sketched, that D exceeds C?

Suppose that Rawls had an argument that C exceeds D. This argument would show that—in response to others' decision not to affirm their sense of justice—it is still better to be just come what may than to be the kind of person who decides case-by-case. I noted earlier that the value of C must be reckoned within the thin theory of the good. That means that the argument showing that it is better to maintain the sense of justice even when others do not would be an argument that shows that the value of expressing one's nature, avoiding hypocrisy and deception, living as friends with others and participating in associations that draw forth talents, all outweigh what could be gained by deciding whether to be just case-by-case, even if others do not maintain their sense of justice. The availability of the last three of these goods—living without hypocrisy and deception, living as friends with others and participating in the right kind of associations—depend upon how we need to respond to others and upon how others treat us. If others do not affirm their desire to be just, then the last two goods will not be available, since—for reasons we saw in Chapter IV—the relevant kinds of friendship and associations depend upon everyone's treating the principles of justice as supremely regulative. Moreover, if others do not affirm their desire to be just, then Player 1 will not need to pretend she has a sense of justice in order to get along with them. In the case where others are not just, then, the first good—the good of expressing one's nature—must be what makes C more highly valued than D. Now recall that the value of A, like the value of C, is reckoned entirely in terms of the four goods. When others are just, the just person enjoys all four goods and not just the one she enjoys when others are not just. So it seems clear that A exceeds C, and that mutual cooperation is to be preferred to affirming one's desire to act from the principles when others do not. Since we are assuming for the moment that C exceeds D, A exceeds D as well. Each player prefers mutual cooperation to a state in which no one affirms her sense of justice. This line of thought shows that if Rawls had an argument that C exceeds D, again all that he would

have to show to demonstrate that mutual cooperation is the preferred outcome is that A exceeds B.

Thus, Rawls is silent about the right column because, regardless of whether D exceeds C or C exceeds D, the comparison in the left column is the one that really needs to be made to show that mutual cooperation is the preferred outcome. Why, then, have I implied that Rawls wants to show members of the WOS face Table II.3, with A > B > D > C? Why have I not left open the possibility that he thinks C exceeds D or that he thinks members of the WOS would be indifferent between the two?

If Rawls's arguments that A exceeds B are successful, then the argument showing that C exceeds D would be an argument that being a just person is Joan's dominant strategy: it would show that Joan should be a just person regardless of what others do. Intuitively, we may think that people should be just regardless of what others do. We may also think that a philosopher like Rawls, who develops a Kantian view, would agree. But showing that justice is each person's dominant strategy—by showing that C exceeds D—would require a very powerful argument. Once we see how the value of C is reckoned, we may wonder whether so powerful an argument is available. As I argued a moment ago, the only good available to the just person when others decide case-by-case is the good of expressing her nature. Members of the WOS would have to value this good extremely highly to value it above what they can gain responding to others by living as they do.

In Chapter VII, we shall ask whether the good of living as a free and equal rational being can do that much work. The congruence argument I am considering in this chapter leaves this good out of account. It asks whether the goods of sincerity, friendship, and association tip Joan's balance of reasons toward affirming her sense of justice. And so it invites us to reckon the values of C in terms of these goods alone. Since they would not be available when others decide case-by-case, C has no value at all in that case. If we assume that there is something to be gained by deciding case-by-case when others do, then—for purposes of seeing how the *Argument from Love and Justice* goes—we can safely assume that D exceeds C, and that Rawls is trying to show that Joan faces Table II.3. Furthermore, as I indicated in §II.2, Rawls does not need so powerful an argument to establish congruence and stability. *All* he needs to show is that A > B > D > C. In that case, as Table II.3 shows, Joan will decide to maintain her desire to act from the principles of justice if she thinks that others will.

Of course, Table II.3 also shows that if *others* do not affirm *their* desire to act from the principles, then it would be rational for Joan not to affirm hers either, and to lead the second kind of life rather than the first. What would be rational for Joan would be rational for everyone else, since Joan is typical. Thus, if each member of the WOS thought that others would opt for the second kind of life rather than the first, they would do so as well. In that case, everyone would lead the second kind of life. The outcome would be the square in the lower half of the right column. What Table II.3 shows, then, is that it is rational for Joan—and every other member of the WOS—to affirm her desire

to act from the principles of justice only if what I called in §II.1 the *mutual assurance problem* is solved. That is why Table II.3 depicts what is referred to as an *Assurance Game*,[3] so-called because when all players face such a table, each will choose to cooperate only if she has the assurance that others will make the same choice. As we shall see, the Rawls of *TJ*—unlike the Rawls of *PL*—thinks that the *mutual assurance problem* is very easily solved in the WOS *once it is shown that A exceeds B*. Indeed, he thinks it is so easily solved that it can be put aside so that he can do the real work of establishing that conclusion. The assumption that "others…have (and…continue to have) an effective sense of justice" is the place where Rawls put it aside.

§VI.3: Conditional Balances and Balance Conditionals

Joan's reasons for maintaining her desire to act from the principles as supremely regulative stem from two sources. They stem, first, from the desires referred to by C_4a, C_4b, C_4c, and C_4d, and second, from the fact that she attaches just as much weight to preserving her sense of justice as she does to satisfying those desires. What we need to know is how the "weights" of these reasons can be totaled, and can be compared to the total weight of the reasons Joan has to live a life in which she trades off her desire to be just case-by-case. The most serious objection to my use of Table II.3 is that it misleads about the clarity of the question. The outcomes in prisoner's dilemma cases are easy to compare because the currency in which the value of outcomes are expressed—jail time—admits of a cardinal measure. The currency in which the value of A and C are expressed does not, and so it is hard to see how the values or "weights" of A and B can be compared.

Throughout *TJ*, Rawls contrasts justice as fairness with teleological theories of justice, and specifically with utilitarianism. He also contrasts it with intuitionism, a view he seems to dispatch early on. One of the problems with both of these rival theories is that they—unlike justice as fairness—are unable to give adequate accounts of Joan's decision. The text and context of the congruence arguments suggest that Rawls is concerned to make this point. Filling in the details of Joan's decision therefore helps to dispel one of the perplexities about Rawls's treatment of congruence, namely, how it fits with the sections that immediately precede it in *TJ*.

Suppose that there is some one good which is the dominant end of human life and that the rational thing to do is to maximize that good. Then Joan could compare the two lives open to her by asking which life maximizes, or does more to maximize, that good when others are just. If we suppose that the dominant end is happiness, then she could compare the two lives by asking which of these two lives is more productive of happiness. Joan will then ask:

3. Douglas G. Baird, Robert H. Gertner, and Randall C. Picker, *Game Theory and the Law* (Cambridge, MA: Harvard University Press, 1994), p. 277.

Will happiness be maximized, or is it more likely to be maximized, if I live a life in which I attain the ends referred to by C_4a, C_4b, C_4c, and C_4d, or if I live a life in which I decide whether to maximize that good case-by-case? If the first life can be shown reliably to be more productive of happiness, then A exceeds B, and Joan knows that it is rational for her to choose that life provided others do. Since Joan is typical, *TJ's Nash Conclusion* would follow:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

Once the *mutual assurance problem* is solved, Rawls could move from C_N to:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

This conclusion gets Rawls to the conclusion he really wants, the *Congruence Conclusion*, which says that members of the WOS would maintain their desire to act from the principles of justice as supremely regulative when they draw up their plans using full deliberative rationality.

I supposed that there is a dominant end because, given what Rawls says about the "symptomatic drift" of teleological theories (*TJ*, p. 560/490), this seems to be the way Rawls would think teleological theory would fill in the details of Joan's decision. Filling in the details this way would raise serious difficulties of intrapersonal comparison, since Joan would have to have some way of comparing how happy she would be leading very different lives. But the real problem with filling in the details this way is that there is *no* dominant end that it is rational to affirm. Members of the WOS have final ends that are "always plural in number" (*TJ*, p. 563/493), including the ends singled out by C_4a, C_4b, C_4c, and C_4d. Recognizing this, Joan will have to face the question of how plans that contain a multiplicity of ends are to be unified. In Chapter VII, we shall see that one of the reasons she would treat her sense of justice as supremely regulative—a reason ultimately connected with C_4a—is that doing so unifies her pursuit of final ends in the only acceptable way. For now, note that teleological theory fills in the details of Joan's choice inaccurately, by implausibly denying that final ends are multiple and supposing, instead, that there is some one currency in which—like jail time in prisoners' dilemma tables—the values of A and B can be computed. Teleological theory cannot, therefore, provide an argument for C_N. I believe one reason *TJ*'s discussions of dominant ends, hedonism, and the unity of the self are placed where they are—immediately before the arguments for congruence—is to make these points.

The intuitionist's way of filling in the details of Joan's decision *does* reflect the plurality of human ends and grants that the ends referred to by C_4a, C_4b,

C_4c, and C_4d are final. But it denies that there is any principled way to balance the value of these ends against the goods available to someone who decides whether to be just case-by-case. Joan would have to discern the balance between A and B that seems right to her and that "seeming" would, as it were, be a brute fact. Joan would therefore have to reach a conclusion about which way her balance of reasons tips without the support of reasons she could make plain to others. Her preference for maintaining her desire to act from the principles of justice as supremely regulative would be—like the preference for a sense of right on Ross's intuitionist account—"without *apparent* reason; it [would] resemble[] a preference for tea rather than coffee" (*TJ*, p. 478/418, emphasis added). This way of discerning the balance between A and B therefore opens what we might call a "justification gap."

Recall that Rawls wants to establish C_N to show C_6 and the *Congruence Conclusion*, and hence to show that justice as fairness would not be destabilized by collective action problems. The existence of the justification gap makes it very difficult to see how that can be shown even if C_N is true. To move from C_N to the other two conclusions, and to avoid a generalized prisoner's dilemma, Rawls must solve the *mutual assurance problem*. To solve that problem, it is not enough to show that C_N is true. C_N must also be generally believed, and it will not be generally believed in the presence of the justification gap. For in the presence of justification gap, *everyone's* judgment that A exceeds B is made on the basis of brute "seemings." Hence *everyone's* preference for treating his desire to be just as supremely regulative is "without apparent reason." But if no one has reasons for preferring A to B that he can make plain to others, then no one has reasons for believing that everyone else prefers maintaining his supremely regulative desire to be just. For all each has reason to believe, some significant number of members of the WOS judge that B exceeds A, and that it is in their interest to take advantage of the justice of others—shirking on their taxes, hiding their gains and passing them along to their children, and otherwise free-riding—when it seems advantageous. In that case, people will respond by being that kind of person themselves. Thus, the justification gap that is opened by intuitionism would destabilize justice as fairness.

Rawls is well aware that he needs to avoid the problem that besets intuitionism. He introduces the question of how Joan is to judge the balance of her motives in such a way as to remind readers of this fact. He says:

> The question now arises whether [Joan's reasons for treating her sense of justice as regulative] are decisive. Here we confront the familiar difficulty of the balance of motives *which in many ways is similar to a balance of first principles*. (*TJ*, p. 572/501, emphasis added)

The problem of how to balance a plurality of first principles is precisely the problem to which intuitionism provides an answer (see *TJ*, pp. 37ff/32ff). Thus, Rawls's wording of the problem faced by Joan is an *ex*plicit reminder that she faces the intuitionist's problem; it is an *im*plicit reminder that Rawls

knows he cannot offer the intuitionist's solution. These reminders confirm a point I have suggested and to which I shall return in §VII.6: Rawls does not consider intuitionism early in *TJ* only to put it aside. His concern to offer an alternative to intuitionism runs throughout *TJ* and many passages are most accurately read as revealing that concern.

Recall that in his first discussion of intuitionism in *TJ*, Rawls says that "the only way therefore to dispute intuitionism is to set forth the recognizably ethical criteria that account for the weights which, in our considered judgments, we think appropriate to give to the plurality of principles" (*TJ*, p. 39/35). If the decision facing Joan really is like the problem of balancing first principles, then this remark suggests that she should determine the weights or values attached to the lives between which she must choose by recourse to some "recognizably ethical criteria." Rawls's critique of teleological theories, and of dominant-end theories in particular, shows that they cannot provide those criteria. For the critique shows that Joan cannot assume a dominant-end theory of the good and attach cardinal payoffs to her alternatives, since those critiques imply that there is no "interpersonal currency" in terms of which such payoffs could be expressed (*TJ*, p. 559/490). We have already seen that those criteria cannot be derived from the full theory of the good without reducing Joan's choice to triviality.

If Joan judges that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as supremely regulative when others do, it will be because of the value or weight she attaches to a life in which she satisfies the desires referred to by C_4a, C_4b, C_4c, and C_4d. Satisfying those desires is, as I remarked at the end of Chapter IV, a shared, partial conception of the good—albeit a thin one. Given the content of that conception, it is not too much of a stretch to describe the conception as in some sense ethical, even if it is not a conception that includes considerations of justice. If this is right, then Joan's choice to satisfy that conception—and hence her choice among lives—are choices that recognizably, or at least discernibly, ethical.

But how is Joan to judge which way her balance of reasons tilts, if not teleologically or intuitionistically? Rawls's answer is that she can tell whether her thin reasons to be just are decisive by making her balance conditional on *another* balance: the balance of reasons in the world as it is. Let us look at what he says.

After noting the similarity between balancing reasons and balancing first principles in the remark I quoted just above, Rawls continues:

> Sometimes the answer is found by comparing one balance of reasons with another, for surely if the first balance clearly favors one course of action then the second will also, should its reasons supporting the first alternative be stronger and its reasons supporting the second alternative be weaker. (*TJ*, p. 572/501)

This is not an easy remark to interpret and Rawls's methodological remarks are far too compressed. To see what he means, recall that I said Rawls's strategy for

establishing *TJ's Nash Claim* exploits what I called the *diversity of descriptions*: it exploits the facts that "the sense of justice…satisfies other descriptions which connect it with reasons specified by the thin theory of the good" and that "the theory of justice [i.e. justice as fairness] supplies other descriptions of what the sense of justice is a desire for" (*TJ*, p. 569/499). Very roughly, Rawls will argue that if Joan's balance of reasons in the actual world would favor replying to the justice of others by maintaining her sense of justice *under a description that connects the sense of justice with reasons specified by the thin theory*, then her balance would favor maintaining it under that description in the WOS, again when others maintain theirs. It follows that in the WOS, her balance tilts in favor of maintaining her desire to act from the principles of justice when others maintain theirs.

The best way to explain this strategy is to begin with some suppositions. Only afterward will we see how those suppositions are justified. The suppositions refer to the four conclusions established in Chapter IV, the conclusions that show Joan's thin reasons to be just. Let us recall those conclusions:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

• Suppose that in the world as it is—the world occupied by us, Rawls's readers—the goods referred to by C_4a, C_4b, C_4c, and C_4d are generally recognized as very great goods. How great? Suppose that they are generally recognized as great enough to outweigh what could be gained and avoided—in the world as it is—by being the sort of person Joan is thinking about being, the sort who replies to others' justice by weighing her desire to be just against her other desires each time temptation presents itself.

• The conditions of the WOS are more favorable than the conditions in the world as it is, in at least this respect: the WOS is a world of perfect compliance. Everyone in the WOS complies with the principles of justice. Suppose that, because of this difference, the following conditional—which I shall refer to as a *Balance Conditional*—holds:

> If, in the world as it is, the goods referred to by C_4a, C_4b, C_4c and C_4d tilt the balance of reasons in favor replying to the justice of others by maintaining one's desire to act from the principles, then they tilt the balance that way in the WOS.

By the first supposition, the antecedent of the *Balance Conditional* is satisfied. Since the second supposition says that the *Balance Conditional* is true, it

follows that in the WOS, Joan's balance of reasons favors replying to the justice of everyone else by maintaining her desire to act from the principles.

How can these goods tip the balance toward maintaining a sense of justice? We saw in Chapter V that Rawls thinks the sense of justice can be described as, for example, "the desire to express our nature" and "the desire to live by the commonly accepted morality of the WOS"; these are the descriptions that "connect [the sense of justice] with reasons supplied by the thin theory." It is because these connections hold that Joan can best or only attain the goods referred to by C_4a, C_4b, C_4c, and C_4d by maintaining her desire to act from the principles of justice when others do. If those goods are as great I have supposed, then it follows that the weight Joan attaches to them—namely, A in Table II.3—exceeds B. In that case, Joan's balance of reasons tilts in favor of affirming her desire to act from the principles in the WOS when others affirm theirs. Since Joan knows what way her balance tilts, it follows that:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

Rawls can then move to C_6 and the *Congruence Conclusion*. This is the way Rawls establishes those conclusions in the *Argument from Love and Justice* and—with some important qualifications—the *Kantian Congruence Argument*.

This reading shows how Joan can judge that her balance favors affirming her sense of justice without opening a "justification gap." For if the *Balance Conditional* is clearly true and if, as I supposed, it is generally recognized that the antecedent of the *Conditional* is true, then Joan's preference for being a just person will not "lack apparent reason." Everyone will recognize that she has good reason for her preference and, as we shall see, she will recognize that everyone else has good reasons for the same preference. Everyone will be in a position to know that everyone else faces Table II.3 and prefers A to B.

We shall see that the two arguments Rawls offers for C_N rely on two different *Balance Conditionals*. The first argument—the *Argument from Love and Justice*—relies on a *Conditional* that refers to C_4b, C_4c, and C_4d. The *Kantian Congruence Argument* relies on a *Conditional* that refers to C_4a. Of course, Rawls's strategy for establishing congruence will succeed only if he can establish the relevant *Balance Conditionals* and the truth of their antecedents. To see how he does so, and to make this very abstract description of his strategy more concrete, we need to turn to the arguments for C_N that Rawls actually offers.

§VI.4: The *Argument from Love and Justice*

After asking whether Joan's reasons to maintain her sense of justice are decisive, Rawls offers three arguments, the first of which is a preliminary argument

that is not intended to establish C_N. Instead, it is supposed to show the weaker conclusion that "however improbable the congruence of the right and the good in justice as fairness, it is surely more probable than on the utilitarian view." This is the *Argument for Relative Stability*. The argument is straightforward and I shall not analyze it here. The second argument, which I shall refer to as the *Argument from Love and Justice*, is supposed to establish C_N and is far more complex and interesting than the first. Rawls introduces the second argument by saying that it suggests a "somewhat different point" than the first. As we shall see in Chapter VII, he later introduces the third argument, the *Kantian Congruence Argument*, by saying that it appeals to considerations that "strengthen[]" the conclusion of the second. The "somewhat" in the introduction to the second argument, and the remarks about strengthening in the introduction to the third, hint at connections among the three congruence arguments that are generally overlooked. Those hints are right.

The *Argument from Relative Stability* shows that congruence is unlikely on the utilitarian view because anyone affirming the principle of utility would find the principle difficult to honor. "It is likely both to exceed his capacity for sympathy and be hazardous to his freedom" (*TJ*, p. 573/500). The case for the relative stability of justice as fairness therefore turns on the question of what commitments citizens of a WOS might later have cause to regret. The *Argument from Love and Justice* and the *Kantian Congruence Argument*, which unlike the first argument *are* supposed to support C_N, turn on the same question.

One problem my reading confronts immediately is that the *Argument from Love and Justice* does not seem to be an argument for:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

It seems to be an argument for a somewhat weaker conclusion. For Rawls introduces the *Argument from Love and Justice* by saying "A somewhat different point is suggested by the following doubt: namely, that *while the decision to preserve our sense of justice may appear rational*, we may in the end suffer very great loss or even be ruined by it" (*TJ*, p. 573/502, emphasis added). The third argument confronts the very same worry. Thus, the overlap or connection between the first argument and the second and third is that the second and the third arguments, like the first, consider what Rawls elsewhere calls "the strains of commitment"—in this case, the strains of a commitment to preserving one's desire to act from the principles (*TJ*, p. 145/126). This point is worth mentioning, since some readers have said emphatically that the only argument that concerns the strains is the *Argument from Relative Stability*.[4]

4. This seems to be the view of Barry, "Search for Stability," p. 886.

More important for present purposes is this. The italicized portion of Rawls's introductory remark suggests the *Argument from Love and Justice* supports the conditional conclusion that *if* Joan judges that her "decision to preserve [her] sense of justice . . . appear[s] rational"—if, that is, she judges that her thin reasons tell in favor of preserving her desire to act from the principles of justice, then she will not regret doing so. But the argument is supposed to *show* that Joan's thin reasons for maintaining that desire are decisive. It is not supposed to show what follows from the assumption that they are.

Let me begin to address this problem—and to show that the *Argument from Love and Justice* does indeed support C_N despite appearances to the contrary—by saying something about the regrets that Joan thinks she might later have about affirming her desire to be just. In *TJ*, §78 Rawls argues that she will not come to regard her sense of justice as a "neurotic compulsion" or the internalization of arbitrary authority. The arguments of that section are therefore intended to rule out one source of regret, but we can imagine many others. We have seen that in the WOS, Joan can gain the goods and relationships referred to by C_4b, C_4c, and C_4d—she can avoid hypocrisy, live with others as friends, and take part in a social union of social unions—only by maintaining a supremely regulative desire to act from the principles of justice. So Joan knows that if she decides not to treat her desire to be just as supremely regulative, she will have to do without these goods and relationships. But she also knows that if she does preserve her desire to be just as supremely regulative, then the relationships which will be open to her will leave her liable to loss or ruin, for her sense of justice may lead her to make great sacrifices for the friends and the institutions to which she is attached. She knows that she can avoid these sacrifices if she becomes the kind of person who weighs her desire to be just against the aversion to loss case-by-case. Is Joan likely to regret that she has chosen a life with attachments that leave her liable to loss? Is she likely to wish that she had chosen a life without those attachments but without the liability to losses either?

Rawls's answer is, of course, "no." Joan knows that if she preserves a sense of justice when others are just, then, at each later point, she will judge—even from within the thin theory—that her balance of reasons favors the kind of life she has chosen to live. This shows that she knows she will "never blame [her]self" for having affirmed her desire to act from the principle of justice "no matter how [her] plans finally work out" (*TJ*, p. 422/371).

As I have implied, Rawls's congruence arguments go by way of *Balance Conditionals*. I shall look at the *Balance Conditionals* at the heart of the *Argument from Love and Justice* in the next section. But the most crucial claim in that argument—and the most interesting one—is a claim Rawls hints at in his *Lectures on the History of Moral Philosophy*. The way Rawls establishes this claim seems to confirm that the *Argument from Love and Justice* supports a weaker claim than C_N.

The claim is found in one of Rawls's lectures on Kant, where Rawls observes that our "fundamental character" is "the ordering that determines

the weight of reasons."[5] What Rawls has in mind, I think, is that our most central traits of character affect what we value. The courageous person attaches less weight to danger than the timid one, and the temperate person values the chance to have another drink less than the bibulous person does.

If this is right, then the weight Joan attaches to her reasons will vary with the character she has. If she is a just person, and is committed to remaining just, the goods she could have gained by being an unjust person will have much less weight for her than they would if she had not made that commitment. To put this claim in game-theoretic terms, if Joan is the kind of person whose life is regulated by the principles of justice, the payoff of the unjust life seems less to her than it would if she were not such a person. Rawls expresses this point especially forcefully in his early paper on "The Sense of Justice," where he says that "the acceptance of the principles of justice implies, failing special explanation, an avoidance of their violation and a *recognition that advantages gained in conflict with them are without value*."[6] So if Joan maintains her desire to act from the principles, then anytime she revisits that commitment in the future, she will judge that her balance of reasons supports it. Whenever Joan asks herself whether she should be just person, she knows that the commitment to being just is not one she will later regret. Of course, the judgments Joan knows she would reach about her balance of reasons are judgments rendered from within the thin theory of the good. The claim I have said is crucial cannot depend upon Joan's taking account of her desire to be just as such. The payoff of the just life, as experienced by the just person, cannot depend upon the fact that it satisfies that desire under *that* description. And it does not. Instead, the payoff depends upon the fact that a just life satisfies that desire under the diversity of *other* descriptions provided by contract theory.

As we have seen, the life of the just person—unlike that of the person who will not commit to justice—is a life in which Joan can realize the goods referred to by C_4a, C_4b, C_4c, and C_4d. The good referred to by C_4a, the good of expressing one's nature as free and equal, is important to the *Kantian Congruence Argument* that I shall look at in Chapter VII. The other goods are what are said to tip Joan's balance of reasons in the *Argument for Love and Justice*. For friendship—the sustained activity of living as friends with others—makes Joan the kind of person who then takes her reasons to sacrifice herself for friends and social forms as stronger than her reasons not to. Joan knows, then, that if she commits to being and remaining just, she will then have open to her the kind of relationships that she will value for their own sake, and that these relationships will shape her so that she discounts what she could gain in the other kind of life. So she knows that if she commits to being just, she will come to value the goods referred to by C_4b, C_4c, and C_4d over B in Table II.3, and the commitment to justice will not be one that she will later regret.

5. Rawls, *Lectures on the History of Moral Philosophy*, pp. 305–6.
6. Rawls, "Sense of Justice," *Collected Papers*, p. 106 (emphasis added).

A comparison may help us to see Joan's situation more clearly. When Joan asks about whether to commit to being a just person, her situation is like that of Jan who is in love and must decide whether to commit to a life-long partnership. Jan may know that the commitment, and the work she will do to maintain it, will change her structure of motives so that she will come to value the goods available in the relationship above what she could get if she left herself free. If so, then she knows that the "strains of commitment" will not be too much to bear and that the commitment is not one she will regret later. Jan may aspire to live up to the ideal of fidelity, and some of the goods of the relationship may be connected with the value she attaches to living up to that ideal. But there may be other goods Jan desires that are available only in the relationship, such as companionship and the good of being loved exclusively by another, that Jan values independently of the value she attaches to fidelity as such. Because these goods are not the object of Jan's ideal-dependent desires, they are—in the relevant way—like the goods Joan values insofar as she follows the thin theory. Suppose Jan knows that, as a result of being in a committed relationship, she will come to value *those* goods more highly than she will value what she could have if she did not commit. Then, at the time she must decide, Jan knows that she will never regret her choice because those goods—analogous to Joan's thin reasons—will always be enough to tip the balance in favor of commitment.

But if we can now see Joan's situation more clearly, we can now see that there are two problems with reading the *Argument from Love and Justice* as an argument for C_N, and not just the one I originally identified.

First, we saw in Chapter V that Rawls thinks members of the WOS have thin reasons to preserve their sense of justice, but this does not itself show that "the decision to preserve our sense of justice" is rational all things considered, for members of the WOS may have reasons not to preserve their sense of justice as well. We would expect that the arguments showing that thin reasons are decisive would show how those countervailing reasons are defeated. The *Argument from Love and Justice* seems only to show how some such reasons are defeated: reasons connected with the possibility of regret. And it seems to depend upon the assumption that other countervailing reasons have already been removed or evacuated of their force. But this is what still needs to be shown.

Second, until it is shown, even the conditional claim that the *Argument from Love and Justice* seems to establish is of questionable significance. For while it may be that if members of the WOS maintain their sense of justice, then they will become the kind of persons who will not regret it, it may also be that if they decided to be the kind of persons who make up their minds whether to be just case-by-case, they will then become persons who would value the goods of *that* life above the goods available to the just person. They might, in short, become the kind of persons who would judge that B in Table II.3 exceeds the payoff of committing to justice when others do the same. If this is so, then at the time she is considering whether to affirm her sense of desire to act from the principles, Joan knows that she will not regret her choice,

whatever she decides. Why, then, does it matter that she would not regret a commitment to justice?

It is important that Joan—like Jan, who must decide whether to commit to a partnership—does not make her choice from some point of view outside both of the lives she is considering. I supposed that Jan is already in love. Joan, like the other members of the WOS is "assume[d] . . . already [to] have" a sense of justice (*TJ*, p. 568/498). The question Joan and Jan ask themselves is therefore whether they should continue leading the kinds of lives they are already leading, provided the relevant other(s) will do the same. If Jan is already in love with someone who reciprocates, then she is already leading a life in which many of the goods of a loving relationship are available. If Joan already has a sense of justice, then she is already living a life in which she enjoys and values the goods of friendship and association. That means that she has already become—or started to become—the kind of person who recognizes that "advantages gained in conflict with [the principles of justice] are without value"[7] and who judges that the value of the goods referred to by C_4b, C_4c, and C_4d exceeds B in Table II.3. To show that someone who has a sense of justice would not regret maintaining it is not to show something about the person Joan could become. It is to show a fact about the person Joan already is. The fact that she would not regret choosing a different life if she altered herself so that she lived *it* is not a fact that moves her, given what she already values.

In the *Tractatus*, Wittgenstein says that the effect of the good will is to make the world "altogether different." "The world of the happy man," he continues "is a different one from that of the unhappy man."[8] Similarly, Rawls might say, the possession of a good will—in the form of a will to act from the principles of justice—makes the world of the just person "altogether different" from the world of the person who lacks that sentiment. "The world of the [just] man" is one that Joan already inhabits in virtue of living in a WOS. And so, contrary to what is suggested by Rawls's introductory remark, the *Argument from Love and Justice* is not just supposed to establish a claim about how Joan would judge her balance of reasons if "the decision to preserve [her] sense of justice . . . appear[ed] rational" on some other grounds (*TJ*, p. 573/502). It is supposed to establish a claim about how Joan would judge her balance of reasons at the time she asks whether to maintain her sense of justice. The claims the argument is supposed to establish are C_N, C_6, and the *Congruence Conclusion*.

§VI.5: Love's Balance

The critical claim in the *Argument from Love and Justice* is that a sense of justice is transformative in at least this sense. Someone who has a sense of justice

7. Rawls, "Sense of Justice," *Collected Papers*, p. 106 (emphasis added).
8. Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*, 6.43.

attaches a different weight to her thin reasons to be just than does someone who lacks the sentiment. I have not yet located that claim in Rawls's text. Moreover, the just person's judgment that her thin reasons to be just outweigh the reasons for leading a different kind of life presupposes that the two sets of reasons can be compared. I have said that the comparison depends upon *Balance Conditionals*. But I have not yet located those *Conditionals* in the text either. The *Argument from Love and Justice* is not easy to make out. In this section, I shall fill in the details of the argument.

When I compared Joan to Jan—the person who must decide whether to commit to a partnership—I did so to draw attention to the transformative effects of love, meaning to suggest that they are like the transformative effects of a sense of justice. The comparison helps to show how the *Argument from Love and Justice* goes. For at the heart of the argument is an interesting argument that relationships of love have systematic effects on motives.

According to the analysis on which that argument depends:

- "being…disposed [to take great chances to help each other] belongs to the[] attachments [of friends and lovers] as much as any other inclination" (*TJ*, p. 573/502).
- So "those who love one another, or who acquire strong attachments to persons and to forms of life, at the same time become liable to ruin: their love makes them hostages to misfortune and the injustice of others" (*TJ*, p. 573/502).
- There is no way to avoid the vulnerability by, as it were, holding back, for "there is no such thing as loving while being ready to consider whether to love, just like that" (*TJ*, p. 573/502).
- So there is no getting around the fact that "once we love we are vulnerable" (*TJ*, p. 573/502).
- "When we love, we accept the possibility of injury and loss" (*TJ*, p. 573/502).
- "Should evils occur," we do not avoid them by ceasing to love those who love us. Rather we treat the evils as "the object of our aversion, and we resist those whose machinations would bring them about" (*TJ*, p. 574/502).

How do these points support the argument Rawls wants to make?

Those who love one another acquire certain attachments and, having acquired them, those who truly love cannot ask whether they should cut their losses and cease loving if loss threatens. Such is the nature of love, even in the world as it is. That is why, even in the world as it is, those who think they have reason to love do not regret their loves; rather they think their balance of reasons tilts in favor of continuing to love. In the WOS, love would leave one less liable to harm than in the world as it is, for treachery and betrayal are absent. "In a society where others are just our loves expose us mainly to the accidents of nature and the contingency of circumstances" (*TJ*, p. 574/502). So if the

balance of reasons tilts in favor of answering love with love in the world as it is, it would surely tilt in favor of doing so in a WOS. Thus Rawls's analysis of love supports the *Balance Conditional*:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of doing the same in the WOS.

We do find ourselves naturally drawn into loving relationships in the world as it is, and we take ourselves to have reason to affirm those loves. And we think Joan would believe the same thing that we do. The antecedent of the *Balance Conditional* is satisfied, so the consequent must be true also. Joan would judge that her reasons tip in favor of answering love with love in the WOS.

But the conclusion Rawls wants to reach—C_N—is a conclusion about reasons to affirm the sense of *justice*, not about reasons to affirm *loves*. The *Balance Conditional* his analysis of love supports is not the one he needs to derive C_N. What he needs is the *Balance Conditional*:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of replying to others' justice by maintaining her desire to act from the principles of justice as supremely regulative in the WOS.

Where does the necessary *Balance Conditional* come from?

Rawls could get from the first *Balance Conditional* to the second, the one he needs, if Joan's affirming that loves are part of her good in the WOS required her to affirm that being just is part of her good there also. And this is just what Rawls seems to think, for when he sums up his analysis of love, Rawls says "if these things are true of love as the world is, or very often is, then a fortiori they would appear to be true of loves in the well-ordered society, *and so of the sense of justice too*" (*TJ*, p. 574/502, emphasis added). The question is why Rawls thinks this last entailment holds.

One possibility is that Rawls thinks Joan cannot love those to whom she wants to be close if she lacks a sense of justice, because wanting to treat someone justly is part of loving her. If she is not committed to being just to those she says she loves, she does not really love them after all. Indeed, we might think, this connection between love and a sense of justice is part of why love makes Joan vulnerable, for only if she has a sense of justice can she recognize—and be harmed by—some of the evils befalling her intimates and some of the evils done her by those she loves. The problem with this reading of Rawls's argument is that it is hard to see why maintaining intimate loves in the WOS would require Joan to have a desire to treat everyone justly, as a sense of justice requires, rather than a desire to treat her intimates justly. Clearly a different reading is called for.

Recall that we have to imagine Joan wondering whether, since justice can leave her liable to ruin, she would be better off not affirming her desire to be

just and doing without the relationships that are open only to the just person. Those relationships embrace those persons and institutions that benefit her, including those to whom she stands in the ties of friendship referred to by C_4c and in the associations referred to be C_4d. In the WOS, the range of persons and institutions that benefit Joan—and that she knows benefit her—is very wide. It includes the just institutions of the WOS and the just people who sustain those institutions. So the relationships referred to by C_4c and C_4d encompass a large part of the WOS. Speaking of the question of whether she should do without *these* relationships, Rawls says "The question is on a par with the hazards of love; indeed, it is simply a special case" (*TJ*, p. 573/502). So I think we have to take the attachments referred to by C_4c and C_4d as among the attachments of love to which Rawls is referring. If in the WOS, Joan would judge that the goods of responding to love with love exceed the goods available by responding otherwise, then she would judge that the goods of answering justice with justice exceeds the goods of responding to others' justice by deciding whether to be just case-by-case. She would, that is, judge that the payoff of answering justice with justice exceeds the value of B in Table II.3.

The analysis of love is therefore supposed to support a *Balance Conditional* that says:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of committing to her loves—including the wide-ranging attachments referred to by C_4c and C_4d—in the WOS.

This is not yet the *Balance Conditional* Rawls needs, but he can get to the one needs with this one in hand. As we saw in Chapter V, Rawls draws on the *diversity of descriptions* to show that Joan can take part in the attachments mentioned in the consequent of this conditional only by treating her desire to act from the principles of justice as supremely regulative. If she does not want to be just to others and sincere toward them, and if she is not committed to taking the principles as supremely regulative, she does not really participate in civic friendship with other just persons or in a social union of social unions. So Rawls can move from the *Balance Conditional* he has to the *Balance Conditional* he needs:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of replying to others' justice by maintaining her desire to act from the principles of justice as supremely regulative in the WOS.

The necessary *Balance Conditional*, together with the claim that Joan would have reason to love in the world as it is, imply that she would judge that the value of the goods referred to by C_4b, C_4c, and C_4d exceeds the value of B in Table II.3, and therefore that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as supremely regulative

when others are just as well. And since Joan poses the real problem of congruence, what I called *TJ's Nash Claim*—follows:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

At the beginning of the previous section, I said that the *Argument from Love and Justice* would show that the desires referred to by C_4b, C_4c, and C_4d provide Joan reasons for maintaining her desire to be just are not defeated by the strains of commitment. The argument for this conclusion depends upon the transformative effects of love, since Rawls assumes that love is transformative in order to establish the first *Balance Conditional*. Because the sense of justice is a form of love, the argument for that conclusion appeals to the transformative effects of justice as well, as I indicated that it would at the end of the last section.

The remark with which Rawls closes the *Argument from Love and Justice* sums up the reasoning that runs from the first *Balance Conditional* to *TJ's Nash Claim*. He says: "taking as a bench mark the balance of reasons that leads us to affirm our loves as things are, it seems that we should be ready once we come of age to maintain our sense of justice in the more favorable conditions of a just society" (*TJ*, pp. 574/502–3). But to say someone is "*ready*" to maintain her sense of justice does not mean that she *will* maintain it. It means that she will maintain it in the right conditions. To see what those conditions are, we need to look again at just what is shown by establishing C_N.

A successful argument for C_N shows that each member of the WOS judges that maintaining his desire to act from the principles is the best response when others maintain theirs. Showing this does not, however, itself show that everyone will affirm her desire to act from the principles. To see this, suppose that though C_N is true, it is not generally thought to be true. Suppose, rather, that Joan is unsure whether the others—or sufficiently many others—judge their balances of reasons as she does. Suppose, that is, that she thinks others—or sufficiently many of them—may prefer B in Table II.3, what they could get by being the kind of person who is ready to act unjustly, to the goods of friendship and association.

Suppose, finally, that while everyone professes be just, Joan distrusts others because she thinks they—or sufficiently many of them—were willing to bear the psychic costs of hypocrisy in order to take advantage of her goodwill. We saw earlier that acting from Rawls's principles of justice is not Joan's dominant strategy. Her balance of reasons tips in favor of affirming her desire to act from the principles when others do the same, but not when they do not or when she thinks they do not. If Joan does not trust others to maintain their desire to act from the principles, she will not maintain hers, simply to protect herself. If others are in Joan's position, they will not maintain their sense of

justice either. The result will be state of mutual *non*cooperation, which is an equilibrium state in Table II.3. Thus when no one has the assurance that others value civic friendship, association, and psychic integrity highly enough to answer justice with justice, justice as fairness will be destabilized. It will be destabilized even if C_N is true.

The reason that establishing C_N is not enough to show stability is that establishing it is not enough to solve the *mutual assurance problem*. Even if C_N is true, each person must still believe that others will maintain their desire to act from the principles or she will not do so herself. But if C_N is true and if everyone is known to have a sense of justice in the first place, then this mutual knowledge is not only necessary, it is also sufficient. For as we saw in §VI.2, Rawls thinks the sense of justice is a sentiment that can only be changed slowly (*TJ*, p. 568/498). Even if circumstances are such that someone who has it would be better off without it, she is sure to be disadvantaged during her transition. Moreover, continuing to live just lives makes great goods available. Those who have a sense of justice would therefore prefer to maintain it. They need *only* the assurance that they will not be taken advantage of if they do.

Thus, the common recognition of others' justice and of the truth of C_N is what is supposed to solve the *mutual assurance problem* and remove or significantly weaken the temptation to preemptive defection. When each knows that everyone has sufficient reason to honor the demands of justice, even if judged from within the thin theory, "it is rational (as defined by the thin theory of the good) for members of the well-ordered society to affirm their sense of justice as regulative of their plan of life" (*TJ*, p. 568/497). This is the claim that I expressed more precisely as:

> C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

We have seen that Rawls can move from C_6 to the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In Chapter III, we saw that there is a trivial route to this conclusion; it should now be clear that the route that goes via C_N secures the nontrivial one.

An important passage in §76 of *TJ* suggests that this is a route Rawls intends to travel. He says that "to insure stability men must have a sense of justice or a concern for those who would be disadvantaged by their defection, preferably both" (*TJ*, p. 497/435). I read the "concern for those who would be disadvantaged by their defection" as a reference to the friendship that members of the WOS must value if the *Argument from Love and Justice* succeeds and if C_N is true. But as we have just seen, instability can be avoided only if

members of the WOS are also *known* to have a sense of justice, and only if C_N is also *known* to be true. Rawls recognizes as much, for he continues "and since *each recognizes* that these sentiments [i.e. a sense of justice and a concern for others] are prevalent and effective"—since each knows that others are just and recognizes that C_N is true—"there is no reason for anyone to think that he must violate the rules to protect his legitimate interests" (*TJ*, pp. 497–98/435, emphasis added).

But on what grounds can "each recognize" that others value friendship and are concerned that she not be disadvantaged? On what grounds do members of the WOS trust one another not to take advantage of others' cooperation? The Rawls of *TJ* says little about how to answer these questions beyond one enigmatic passage in which he seems to suggest that the *mutual assurance problem* is solved because the truth of C_6 is brought about "by public institutions" (*TJ*, p. 336/296). To see how Rawls thinks the WOS solves the *mutual assurance problem*, it helps to reflect on the ways in which the WOS differs from situations in which solving that problem raises more interesting and difficult questions.

One such situation is that of persons who want to coordinate their behavior but lack a convention for doing so. There may be several conventions—such as everyone driving on the right and everyone driving on the left—that would bring about equally good equilibrium states. The *mutual assurance problem* in these cases is the problem of assuring people who are not cooperating that everyone will adopt the same course of action. Another and very different such situation is that of a Stag Hunt. In Stag Hunts, each can gain more for himself if all cooperate in hunting stags than if each goes his own way and hunts hare, but hunting hare is also an equilibrium state. If all are currently hunting hare, then the *mutual assurance problem* is that of "lifting" everyone from a sub-optimal Nash equilibrium of noncooperation, to the one of mutual cooperation.[9]

The situation of members of the WOS differs from the case of those who want a convention but lack one, and of stag hunters who are all hunting hare, because members of a WOS are already cooperating. Moreover, in situations of the first two kinds, everyone recognizes that no one else has an incentive to "ride free" on the cooperation of others. Driving on the left when others drive on the right is disastrous, so an agreement to drive on the right would be stable. While an agreement to hunt stags may not be stable, the temptation to ride free on the adherence of others is not what destabilizes the agreement, for even the hunter who defects from the stag hunt knows that the most he can expect—namely, the hare—is still less than he could have had if everyone had cooperated, including himself.

A more useful comparison with the WOS would be cases in which each thinks others might believe they have something to gain from defecting from

9. See Bryan Skyrms, *The Stag Hunt and the Evolution of Social Structure* (Cambridge: Cambridge University Press, 1993).

an agreement while others adhere to it. Rawls himself suggests such a case when he observes that disarmament agreements are imperiled by *mutual assurance problems* (*TJ*, p. 336/296). Seeing how the members of the WOS differ from parties to a disarmament agreement can suggest how the *mutual assurance problem* is removed in the WOS.

Arms control agreements—like truces between religious sects—are typically agreements actually reached by parties who have a history of conflict. That there was conflict at all shows that at least one party believed it had something to gain by defeating its opponents. Clearly the fact that an agreement or a truce has been reached does not itself show that the parties who once held this belief have given it up. And so long as each party thinks others hold on to that belief, each will fear that the others think they can gain by taking advantage of the peace brought about by agreement. If parties refrain from preemptive defection, so that the agreement remains stable, the peace that prevails is still what the Rawls of *PL* calls "a modus vivendi" (*PL*, pp. 146–47). What really makes for a modus vivendi, though Rawls never puts it this way, is the fact that the *mutual assurance problem*—though solved at least temporarily—is not solved by parties' trust in one another's good will. The reason it is not solved by mutual trust in disarmament agreements and truces is that the history of conflict makes it difficult for parties to trust one another.

Agreement on principles of justice differs in significant respects from agreement on a cessation of hostilities. For one thing, agreement in the OP is hypothetical. More important for present purposes, adherence to the terms that would be agreed to is not immediately preceded by a period of conflict. There is therefore no recent history of conflict among members of the WOS that provides reasons for mistrust among them. Rather, as I argued in §VI.4, Joan and the other members of the WOS already live in the world of the just person. So when Joan asks whether to maintain her sense of justice, she is asking how best to respond to persons who have a history of treating her with justice and to institutions that have promoted her good and that of people she cares for.

Moreover, the very fact that Joan has a sense of justice indicates that others have acted with "evident intention" to honor their obligations and that she can recognize how she and others have benefited from their institutions, since these are conditions of her developing that sentiment (*TJ*, pp. 490–91/429–30). The evidence of others' intentions, provided by their past behavior, is what gives Joan reason to think that they are just. Their evident intentions also give her reasons to think they care about the good of others, judge that they have less to gain by defecting from the agreement than they do by continuing to honor it, and can be trusted when they profess to be just. And so it is their evident intentions that allow Joan to infer, not only that others are just, but that C_N is true. This solves the *mutual assurance problem* since Joan knows that they have no reason to stop being just. She will therefore commit to maintaining her desire to act from the principles of justice and to treating the principles as supremely regulative. "Being rational for anyone," Rawls says, this decision "is

rational for all" (*TJ*, p. 568/497). The WOS is in equilibrium. Since the reasons Joan has to be just—and the reasons she knows that others have to be just—stem from enduring desires that are constantly reinforced by just institutions, that equilibrium will be stable in the long run.

Though Rawls assumes strict compliance in the WOS (*TJ*, p. 8/8), he also argues that should someone defect and behave unjustly, he will be moved by "association guilt" to accept just punishment and to "make good the harms caused to others" (*TJ*, p. 470/412). An account of retributive justice would show how punishment and reparation restore the justice of the WOS, and so return it to equilibrium. We saw in Chapter II that Hobbes thought terms of cooperation could be stabilized only by the institution of a sovereign who was known effectively to enforce severe penalties for defection. Contrasting his treatment of stability with Hobbes's—and as if to summarize the arguments I have laid out in this section—Rawls says "now it is evident how relations of friendship and mutual trust, and the public knowledge of a common and normally effective sense of justice, bring about the same result" (*TJ*, p. 497/435). The "relations of friendship and mutual trust," the existence of a "common and normally effective sense of justice," and "public knowledge" of a common sense of justice, all result from the operation of just institutions, institutions which implement the principles of justice. The stability that justice as fairness would enjoy if the *Argument from Love and Justice* is successful is therefore the kind of stability Rawls promised to show: stability that is *inherent* rather than *imposed*.

I shall argue in Chapter VIII that Rawls came to think the argument was not successful, and that a conclusion relevantly like *TJ's Nash Claim* had to be established on other grounds. The failure of the *Argument from Love and Justice* is one of the reasons that he made the turn to political liberalism. I want to close this chapter with some remarks about the significance of the argument.

§VI.6: Four Comments on the Argument

One of the crucial moves in the *Argument from Love and Justice* is the assimilation of relationships referred to by

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

to relationships of love. For it is only by assimilating those relationships to relationships of love that Rawls can move from his analysis of love to the *Balance Conditional* that says:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that

> her balance of reasons tilts in favor of committing to her loves—including the wide-ranging attachments referred to by C_4c and C_4d—in the WOS.

We might wonder what warrants the assimilation, since relationships founded on the giving and receiving of love might seem very different from those founded on the giving and receiving of justice. Rawls does assert a similarity between the sentiment of love and the sense of justice. As he says that "there is no such thing as loving while being ready to consider whether to love, just like that" (*TJ*, p. 573/502), so he says that "a just person is not prepared to do certain things, and if he is tempted too easily, he was prepared after all" (*TJ*, p. 569/498). This suggests that he thinks relations of justice have systematic effects that are like the effects of relations of love—as they would have to have if this last *Balance Conditional* is true.

Some hint that Rawls thinks the sense of justice has such effects can be found in a very different source, Rawls's *Lectures on the History of Moral Philosophy*. In the first of his lectures on Kant, Rawls contrasts Kant's view of talents of mind as "gifts of nature" with a good will. He writes:

> a good will is not a gift. It is something achieved; it results from an act of establishing a character, sometimes by a kind of conversion that endures when strengthened by the cultivation of the virtues and of the ways of thought and feeling that support them.[10]

Rawls's claim that a good will is "achieved" and "strengthened by cultivation" suggests that, just as persons like Joan may decide to affirm and work to maintain their sense of justice, so Kant thinks moral agents decide to affirm and work to maintain their good will. In the Introduction, I called attention to Rawls's choice of the word "conversion" and its religious overtones. Here we need to be attuned to another of its resonances. The choice of this word to describe how a good will can be achieved suggests that Rawls thinks the achievement of a good will transforms one's structure of motives and "ways of thought and feeling." If Rawls also thinks, as I believe he does, that to have a sense of justice is to have—or is an important part of having—a good will, then the quoted passage suggests that maintaining a sense of justice is transformative. The passage therefore goes some way to confirming my interpretation of the *Argument from Love and Justice*.

It is surprising that Rawls does not defend the assimilation of relationships of justice to relationships of love in the course of that argument. I think Rawls would reply that the demand for a defense presupposes a claim he would not be willing to grant: the claim that relationships of justice and relationships of love are significantly different in kind. Rawls gives some indication that he thinks they are *not* significantly different in his remark that "the sense of justice is continuous with the love of mankind" (*TJ*, p. 476/417) and in related passages (e.g., *TJ*, pp. 191–92/167). This passage suggests that we read the crucial move in the

10. Rawls, *Lectures in the History of Moral Philosophy*, p. 155.

Argument from Love and Justice as an instance of what we have seen before. It is a case of Rawls building on groundwork that was laid down much earlier in *TJ* precisely so that he could draw on it later to establish congruence. Unfortunately, Rawls's earlier treatments of love and justice are—like the one I just quoted—brief and cryptic, and so are very difficult to interpret in ways that lend argumentative support to the crucial move in the *Argument from Love and Justice*. Rather than trying to extract an argument for that move from Rawls's earlier remarks, I am inclined to proceed in reverse. I am inclined to grant Rawls at least the prima facie plausibility of the assertion he needs for the *Argument from Love and Justice* and to take the move as giving some indication of what Rawls had in mind when he asserted the continuity between justice and a love of mankind.

The fact that one of Rawls's arguments for congruence—surely among the most important arguments in part III of *TJ*—depends upon the desires referred to by C_4b, C_4c, and C_4d underlines the second point I wish to make about the *Argument from Love and Justice*. To describe justice as fairness as an individualistic conception of justice—as many critics do—is to caricature Rawls's view. It is to distort the subject by exaggerating the presence of one of its features at the expense of others that a more realistic picture would show to be equally prominent. For while the argument for the principles in Part I of *TJ* proceeds from individualistic assumptions, Rawls's treatment of stability—and in particular his treatment of congruence—depends upon the presence of desires for sociability. This would, I believe, be obvious if we looked at the content of Joan's ideal-dependent desires and at the way that congruence and stability depend upon them. But it is clear even when we look at what Joan desires insofar as she follows the thin theory. The congruence arguments that appeal to those desires depend upon the claim that human beings living under just institutions will naturally develop the desire to live among others in certain characteristic ways, as a certain kind of person. The conception of a person who wants to live in those ways is not of a solitary individual, but of a person with wide-ranging loves and attachments that affects his structure of motives and the weights he attaches to lower-order desires. That conception—together with the ideals that justice as fairness includes—helps to make Rawls's conception of justice a very attractive a view. Those who would stress the individualist premises of the argument for the principles will therefore miss some of what makes the view most appealing.[11]

11. In his characteristically penetrating review of Rawls's *Lectures on the History of Political Philosophy*, Colin Bird argues that "the picture of Rawls as a crassly individualist political thinker cannot survive a close reading of these lectures." Bird's argument for this assessment turns on his recognition that Rawls thinks individuals in ordinary life can be moved by desires to conform to reasonable principles for their own sake. The desires Bird points to are what Rawls calls "principle-dependent desires." I mean to offer a further argument against the individualist reading of Rawls, one that turns on the possibility that members of the WOS will be moved by a conception-dependent desire to be the kind of person discussed in the text. Bird's review of Rawls appears in *Ethics* 117 (2007): pp. 784–90.

Third, the thought that justice as fairness is individualist in some objectionable way is abetted by a reading of the congruence arguments according to which Joan is an egoist who needs to be shown that it is good to be just. Rawls denies this reading (*TJ*, p. 568/497). Yet, as I said at the end of §II.3, the feeling may persist that in following the thin theory, Joan is reasoning as an egoist would. It is important to see that she is not.[12]

An egoist, Rawls rightly reminds us, "is someone committed to the point of view of his own interests. His final ends are related to himself: his pleasures and social prestige, and so on" (*TJ*, p. 568/497). The point of view of the thin theory is not the point of view of one's own interests. As we saw in §III.4, the person who follows the thin theory can be moved—even moved to injustice—by desires for ends that are not selfish in any obvious way. She may want to cheat on her taxes so that she has more money to give away, for example. She may even be moved by what she takes to be demands of morality. What she is not moved by are the demands of justice as fairness as such. Moreover, the congruence arguments show that Joan affirms her sense of justice because she knows that she can secure other final ends—the ends of psychological integrity, friendship, and association, and the end of expressing her nature as free and equal—only by taking principles of justice as supremely regulative. The last three of these ends, at least, do not seem to be among the final ends of the egoist. Even in following the thin theory, Joan pursues different ends than the egoist would.

Perhaps if human beings were all egoists, and pursued only the egoist's ends, we would need Hobbes's solution to the problem of stability. According to Hobbes, stability is brought about by a sovereign who transforms the payoffs of cooperation and defection. The result is that members of the Hobbesian society do not face a prisoners' dilemma. Each person sees that it is in his interest to cooperate, but the interests each person has continue to be the interests of the egoist. On Rawls's view, just institutions transform the payoff table Joan faces, not by transforming the payoffs in the manner of a Hobbesian sovereign, but by transforming Joan. They do so by encouraging the pursuit of a number of final ends, including the ends of justice as such, the objects of the ideal-dependent desires, and the objects of the desires referred to by C_4a, C_4b, C_4c, and C_4d. Desires for all these ends stabilize justice as fairness.

The transformation of Joan that makes stability possible is not a transformation of her nature. Recent centuries have made us all too familiar with attempts to transform human nature for political ends, attempts most thoughtfully explored and decried by Isaiah Berlin.[13] Rawls would agree with Berlin that these attempts have been catastrophic. He also thinks they have

12. I am indebted to my colleague James Sterba for showing me the importance of taking up this point.

13. See, for example, Isaiah Berlin, "The Decline of Utopian Ideas in the West," in his *The Crooked Timber of Humanity* (New York: Alfred A. Knopf, 1991), ed. Henry Hardy, pp. 20–48, especially pp. 47–48.

been unnecessary, since a stably just society does not require such a transformation. In *The Law of Peoples* he says he concurs with Rousseau, thinking the task of a theory of justice is to take "*men* as they are" and to identify "*laws*"—and not men—"as they might be."[14] And so he would insist that the final ends Joan is encouraged to pursue are ends that are natural to her in this sense: according to the laws of psychology, human beings will come freely to desire these ends if we live under just institutions. It follows that, in this sense of "natural," our ends are not naturally confined to those of the egoist and we are not the natural egoists Hobbes supposed us to be.

The fact that we are not natural egoists opens the possibility of a non-Hobbesian account of stability. Justice as fairness, when institutionalized, can stabilize itself, in part, by encouraging desires for the natural final ends to which the congruence arguments appeal. Rawls suggests as much in "The Sense of Justice," in a passage which anticipates a contrast with Hobbes that he draws in *TJ*. Speaking of the WOS, he says:

> Thus not only may such a system of cooperation by stable in the sense that when each man thinks the others will do their part there is no tendency for him not to do his; it may be inherently stable in the sense that the persistence of the scheme generates, in accordance with the second psychological law, inclinations which further support it. The effect, then, of relations of friendship and mutual trust is analogous to the role of the sovereign; only in this case it is the consequence of *a certain psychological principle of human nature* in such systems.[15]

Of course, Rawls is not the only thinker to note that the *mutual assurance problem* is solved, and prisoner's dilemmas avoided, among those who trust one another.[16] But if the bearing of friendship and commitment on collective action problems is commonly recognized, thinkers who have recognized it have not seen how the requisite attachment among players can be developed, let alone shown how the terms of cooperation—when institutionalized—could themselves encourage such attachment and sustain it over time. *TJ*'s discussions of the connection between moral and natural attitudes, and of the connection between justice and friendship, have attracted very scholarly little

14. John Rawls, *The Law of Peoples* (Cambridge, MA: Harvard University Press, 2001), p. 7 (emphases added).

15. Rawls, "Sense of Justice," *Collected Papers*, p. 105 (emphasis added).

16. Edna Ullmann-Margalit is especially clear on these points; see her *Emergence of Norms*, p. 21. Robert Axelrod implies them, indicating that one of his findings is significant because it shows how the prisoner's dilemma can be averted among those who are *not* friends; see Axelrod, *Evolution of Cooperation*, p. 87. Amartya Sen has suggested that players avoid prisoners' dilemmas by being committed to one another; see Sen, "Rational Fools," pp. 340–41. Sen has analyzed the notion of commitment with considerable subtlety; see, for example, his "Why Exactly is Commitment Important for Rationality?," *Economics and Philosophy* 21 (2005): pp. 5–13.

attention. This neglect is unfortunate. One of Rawls's tremendous contributions in *TJ*, part III is that of showing how just institutions foster wide-ranging and long-lasting attachments of civic friendship in the WOS, as well as attachments to just institutions. In this chapter, I have tried to show how the *Argument from Love and Justice* draws on the existence of those attachments in a WOS to show that justice as fairness would avert the threat of collective action problems and stabilize itself.

The final point I want to make about the *Argument from Love and Justice* concerns its limitations. The passage I quoted from "The Sense of Justice" says explicitly that the friendship and mutual trust of the WOS depend upon what Rawls calls "the second psychological law." That is the law that governs development of what *TJ* refers to as "the morality of association." The law says:

> given that a person's capacity for fellow feeling has been realized by acquiring attachments in accordance with the first law, and given that a social arrangement is just and publicly known by all to be just, then this person develops ties of friendly feeling and trust toward others in the association as they with evident intention comply with their duties and obligations, and live up to the ideals of their station. (*TJ*, p. 490/429)

According to this law, the development of "friendly feeling and trust toward others" depends upon each person's being able to see that others intend to do their part. That is why evidence of intention is necessary to solve the *mutual assurance problem*, as we saw at the end of the previous section. If the WOS were small enough that each person's intentions were evident to everyone else, then not only might each person's defection do perceptible harm to everyone else, but mutual concern and trust might extend so widely that everyone would have a lively "concern for those who would be disadvantaged by their defection" (*TJ*, p. 497/435). The argument that everyone in the WOS would have a sense of justice, together with the *Argument from Love and Justice*, would suffice to show inherent stability. In that case, we might say—without too much of a stretch—that justice as fairness would be stabilized by the morality of association.

But the WOS would be a large, modern society in which members' intentions will not be evident to everyone else. Rawls cannot count on the friendship that develops in accord with the second psychological law being all-embracing. And so he cannot count on each person's balance of reasons being tipped toward justice by friendship, nor can he count on the *mutual assurance problem* being as easily solved as it would be were the WOS much smaller.[17] The *Argument from Love and Justice* shows some of the reasons members of the WOS would have for thinking that their balance of reasons tips toward answering justice with justice, and some of the grounds on which the *mutual assurance problem* is solved, but it cannot be the whole story.

17. Cf. Ullmann-Margalit, *Emergence of Norms*, pp. 22, 47.

Rawls himself is not unaware to this problem. That is why, even in "The Sense of Justice," he rests his case for stability upon a third psychological law as well.[18] That law governs the development of what Rawls calls "the morality of principles." It states the conditions under which members of the WOS would develop the disposition to act from principles of justice for their own sake. Having argued that they would develop that disposition, Rawls needs to show that members of the WOS would judge, from within the thin theory, that their balance of reasons tips toward maintaining it. And so he needs an additional argument for congruence.

The *Argument from Love and Justice* depends upon the claim that members of the WOS realize certain elements of their nature when they treat their desire to act from the principles of justice as supremely regulative. For it depends upon the claim that by being just when others are just, they are able to satisfy their natural desires for psychological integrity, for friendship, and for participation in associations that draw forth their own and others' talents. The *Two Conjunct Reading* of the Aristotelian Principle—which I introduced in §VI.1 and which stressed that "the exercise of our natural powers is a leading human good" (*TJ*, p. 426, note 20/374, note 20)—suggests what the *Argument from Love and Justice* confirms: that members of the WOS would judge that realizing these elements of their nature belongs to their good. But that argument leaves one important element of our nature out of account. It does not show that by being just, we realize our nature *as free and equal rational beings*. *TJ*'s other congruence argument—the *Kantian Congruence Argument*—purports to show what the *Argument from Love and Justice* cannot: that members of the WOS would affirm the disposition to honor principles of justice for their own sake. It purports to show that by showing that maintaining that disposition is the only way members of the WOS can realize their nature as such beings. In Chapter VII, I take up that argument.

18. Rawls, "Sense of Justice," *Collected Papers*, p. 106.

VII

Kantian Congruence and the Unified Self

In Chapter VI, I laid out what I called Rawls's *Argument from Love and Justice*. In this chapter, I shall take up Rawls's other congruence argument, an argument I shall refer to as the *Kantian Congruence Argument*.[1]

The *Kantian Congruence Argument* as Rawls states it is very difficult to follow, in part because Rawls's exposition of the argument does not seem to follow the sequence of his reasoning. Moreover, though the argument draws on considerations introduced when Rawls lays out Joan's reasons to be just—it draws on what I called in §V.3 the argument from "the desire to express our nature"—it is not clear from the text exactly how that argument fits into the *Kantian Congruence Argument*. Yet for all its obscurity, the *Kantian Congruence Argument* is very important. I said in Chapter VI that Rawls came to think the *Argument from Love and Justice* failed and that its failure was part of what led to his political turn. I do not think, then, that the *Kantian Congruence Argument* is the only part of *TJ*'s treatment of congruence with which Rawls became dissatisfied. It is, however, *one* of the arguments with which he became dissatisfied. If we are to understand his turn to political liberalism, we need to know why he became dissatisfied with it. To see *that*, we need to see exactly how the argument goes. Laying out the *Kantian Congruence Argument*, and distinguishing my reconstruction of the argument from other plausible reconstructions, are the primary tasks of this chapter.

1. Following Samuel Freeman.

The work of piecing together the *Kantian Congruence Argument* does not just shed light on the reason for Rawls's political turn. It also sheds a great deal of light on *TJ* itself by extending what is generally appreciated about Rawls's debt to Kant. Standard defenses of the Kantian Interpretation of justice as fairness quite rightly stress the Kantian argument Rawls provides for the principles of justice.[2] What is less often discussed is Rawls's Kantian conception of the unity of practical reason, and his argument that practical reason is unified when we treat our sense of justice as supremely regulative over the course of life.[3] In §§VII.6 and VII.7, I shall try to show what that argument contributes to the *Kantian Congruence Argument*. The details of the *Kantian Congruence Argument* also shed light on the much-controverted question of whether the OP is essential to justice as fairness. Some of Rawls's defenders have tried to answer criticisms of justice as fairness by showing that it is not. In §VII.9, I argue that it is essential to justice as fairness as laid out in *TJ*, to a limited but precise extent.

§VII.1: An Overview of the *Kantian Congruence Argument*

In this section, I will sketch my reading of the *Kantian Congruence Argument.* It therefore will be helpful to have the text of the argument before us. Rawls lays out the argument immediately after concluding the *Argument from Love and Justice*. He writes:

> One special feature of the desire to express our nature as moral persons strengthens this conclusion. With other inclinations of the self, there is a choice of degree and scope. Our policy of deception and hypocrisy need not be completely systematic; our affective ties to institutions and to other persons can be more or less strong, our participation in the wider life of society more or less full. There is a continuum of possibilities and not an all or nothing decision, although for simplicity I have spoken pretty much in these terms. But the desire to express our nature as free and equal rational beings 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

2. I have in mind Stephen Darwall, "A Defense of the Kantian Interpretation," *Ethics* 86 (1976): pp. 164–70, and Arnold Davidson, "Is Rawls a Kantian?," *Pacific Philosophical Quarterly* 66 (1985): pp. 48–77.

3. An exception is a splendid but, unfortunately, little-cited piece by Thomas Pogge: "The Kantian Interpretation of Justice as Fairness," *Zeitschrift für philosophische Forschung* 35 (1981): pp. 47–65.

> 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. It is a desire to conduct oneself in a certain way above all else, a striving that contains within itself its own priority. Other aims can be achieved by a plan that allows a place for each, since their satisfaction is possible independent of their place in the ordering. But this is not the case with the sense of right and justice. . . . What we cannot do is express our nature by following a plan that views the sense of justice as but one desire to be weighed against others. For this sentiment reveals what the person is, and to compromise it is not to achieve for the self free reign but to give way to the contingencies and accidents of the world. (*TJ*, pp. 574–75/503)

The first sentence of this passage refers to the conclusion of the *Argument from Love and Justice* and says that a "special feature" of the desire referred to in C_4a—"the desire to express our nature as moral persons"—"strengthens" it. We saw in Chapter VI that one of the conclusions of that argument is *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

On my reading, this is the conclusion to be strengthened. The *Kantian Congruence Argument* strengthens the conclusion not—as might be supposed—by supporting a stronger variant of it, but by offering a stronger defense of the conclusion than the *Argument from Love and Justice* does. The *Kantian Congruence Argument* strengthens Rawls's defense of the conclusion by building on and strengthening the argument from C_4a that I laid out in §V.3.

Part of what makes it so difficult to see how the *Kantian Congruence Argument* strengthens the conclusion of the *Argument from Love and Justice*—indeed, part of what makes Rawls's exposition of the *Kantian Congruence Argument* somewhat confusing—is the difficulty of locating the conclusion of the *Kantian Congruence Argument* in Rawls's text. What might seem to be the conclusion is expressed by a sentence that is oddly placed in the middle of the passage: "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." I shall refer to this sentence as the "ostensible conclusion" of the *Kantian Congruence Argument*.

There is a reading of the *Kantian Congruence Argument* according to which the ostensible conclusion is, if not the real conclusion, at least quite

close to it. I shall look at that reading in §VII.3. On my reading, by contrast, the ostensible conclusion is *only* ostensible. As I said in §V.2, I think the conclusion Rawls ultimately wants to reach in his treatment of congruence is what I called the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

On my reading, not only is C_C the conclusion the Rawls of *TJ* ultimately wants to reach, but he thinks the *Kantian Congruence Argument* enables him to reach it.

The *Congruence Conclusion* C_C concerns the viewpoint of full deliberative rationality, and I have said that the congruence arguments concern Joan, who judges from within the thin theory of the good. That means that the *Kantian Congruence Argument*—like the *Argument from Love and Justice*—must reach the *Congruence Conclusion* by way of a conclusion about judgments reached from within the thin theory. That is why I read the *Kantian Congruence Argument*, like the *Argument from Love and Justice*, as supporting the *Congruence Conclusion* by appealing to *TJ's Nash Claim* C_N and to the other important claim about judgments reached from within the thin theory:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

As I shall explain in §VII.3, I take the *Kantian Congruence Argument* to resemble the *Argument from Love and Justice* in another important way. On my reading, it—like the earlier congruence argument—supports *TJ's Nash Claim* by appealing to a *Balance Conditional*.

Thus on my reading, Rawls's statement of the *Kantian Congruence Argument* is highly compressed and elliptical. When fully laid out, it begins with an argument for

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

It builds on and strengthens the argument from C_4a that I laid out in Chapter V to reach what I called its "ostensible conclusion"—the claim that "in order to realize our nature we have no alternative but to plan to preserve our sense of justice as governing our other aims"—a conclusion which, on my reading, concerns a judgment reached by the typical member of the well-ordered society (WOS) following the thin theory. Like the *Argument from Love and Justice*, the *Kantian Congruence Argument* then draws on a *Balance Conditional* to move from the "ostensible conclusion" to *TJ's Nash Claim*, but the strength-

ened argument from C_4a provides a stronger basis for that claim than the *Argument from Love and Justice* did. Like that argument, the *Kantian Congruence Argument* then uses *TJ's Nash Claim* to solve the *mutual assurance problem* and reach C_6. Finally, the argument moves from C_6 to the *Congruence Conclusion* C_C and inherent stability.

This interpretation of the *Kantian Congruence Argument* has a number of strengths.

First, it reads the argument so as to fit with Rawls's stated intention to establish that "it is rational for someone, as defined by the thin theory, to maintain his sense of justice" and to establish that it is rational by showing that "the plan of life which does this is his best reply to the similar plans of his associates" (*TJ*, p. 568/497).

Moreover, by seeing how Rawls establishes this latter claim, which I called *TJ's Nash Claim*, we can see that the *Kantian Congruence Argument* is not an alternative to his "balance of reasons" arguments.[4] Rather, the argument *concerns* each person's balance of reasons. For on my reading, the *Kantian Congruence Argument*, like the *Argument from Love and Justice*, establishes that each person's reasons for maintaining her sense of justice are decisive by showing that the balance of her reasons tips in that direction. Thus, the *Kantian Congruence Argument*—like the *Argument from Love and Justice*—is just the kind of argument Rawls promised to offer (see *TJ*, p. 572/501).

A further advantage of my reading is that it brings to light a number of similarities between the *Kantian Congruence Argument* and the *Argument from Love and Justice*. I have already mentioned several of them. Let me mention another. The *Argument from Love and Justice* responds to Joan's worry that she would regret her decision to lead a certain kind of life, a life regulated by the desire to be just. I read the *Kantian Congruence Argument* as responding to that worry as well. On my reading, the first part of the passage I quoted from *TJ*, pp. 574–75/503 is supposed to establish that leading the life of the just person is the only way to satisfy the desire referred to by C_4a, the desire to express one's nature. Joan makes the decision to lead such a life in the face of temptation to be a different kind of person, one who decides case-by-case whether or not to act justly. She might be tempted by such a life despite thinking that deciding case-by-case would not express her nature as a free being. There is, however, another possibility. Joan's temptation to be the kind of person who decides case-by-case might be heightened by the thought that such a life *does* express her nature as free, precisely because a life without a precommitment to justice leaves her free to decide how to behave as cases arise. The end of the passage—beginning with "What we cannot do"—can seem superfluous or merely hortatory.[5] On my reading, however, that part of

4. As Brian Barry would have it; see Barry, p. 886.

5. Gerald Cohen dismissed passages like it as parts of a "high-pitched homily." G. A. Cohen, "Where the Action Is: On the Site of Distributive Justice," *Philosophy and Public Affairs* 26 (1997): pp. 3–30, p. 17.

the passage has an important function. It is supposed to clinch the *Kantian Congruence Argument* by showing that the second form of life under consideration does not express our nature as free but is, in fact, the choice Joan would regret.

Though my interpretation makes sense of a great deal of the text surrounding the *Kantian Congruence Argument*, it seems to have little basis in the text of the argument itself. What I called the "ostensible conclusion" is found barely halfway through the reconstruction of the argument that I have sketched. It therefore lies at some argumentative distance from the conclusions I have said Rawls really wants to reach, and that distance is not bridged in the text. Indeed, little of the reconstructed argument I sketched can actually be found in the passage I quoted at the beginning of this section.

In my view, however, we should not expect Rawls to lay out the whole of the *Kantian Congruence Argument* in §86, since there as elsewhere he presupposes acquaintance with earlier parts of *TJ*. While that argument's reliance on the earlier argument from C_4a is not immediately evident, I believe that the allusion to C_4a in the opening sentence is a reminder of that argument and signals that the *Kantian Congruence Argument* will rely upon it. Moreover, we have already seen that in the *Argument from Love and Justice*, Rawls devotes himself to establishing *TJ's Nash Claim*, and he treats the solution to the *mutual assurance problem* and the move to the *Congruence Conclusion* as if they went without saying. We should not be surprised that these moves receive similar treatment in the *Kantian Congruence Argument*.[6] I therefore think that we should read the passage I quoted at the beginning of this section as presupposing the argument from C_4a, and as taking what follows the "ostensible conclusion" as steps to be supplied by the reader.

To defend my reading, I shall supply the parts of the argument that Rawls did not, and I shall show how they fit with the argument for the "ostensible conclusion." I shall begin at the beginning, with the argument from C_4a.

§VII.2: The Argument from C_4a

I have already called attention to the opening sentence of the passage in which the *Kantian Congruence Argument* is laid out, where Rawls says that "one special feature of the desire to express our nature as moral persons strengthens" the conclusion of the *Argument from Love and Justice*. The sentence suggests that Rawls thought there was some weakness in the *Argument from Love and*

6. Interestingly, in Chapter X, we shall see that the Rawls of *PL* recognized that he had to give some of the parallel steps—especially the *mutual assurance problem*—considerably more attention.

Justice which the *Kantian Congruence Argument* helps to remedy. What, exactly, is the weakness and why does the argument suffer from it?

The *Argument from Love and Justice* begins with the claim that members of the WOS have reasons to be just that stem from desires they all have, the desires referred to by:

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

and

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

The argument then purports to show that these reasons tell decisively in favor of each person's maintaining her desire to regulate her plans by the principles of justice when others do, so that C_N—*TJ's Nash Claim*—is true. It purports to show that by showing that the desires referred to by C_4b, C_4c, and C_4d can only be satisfied by maintaining the desire to act from the principles as supremely regulative, again when others do.

But beginning in the second sentence of the quoted passage, Rawls concedes that the claim from which he derived C_N is not true. Joan can incorporate the end referred to by C_4b into her plan of life while pursuing a partial and unsystematic policy of deception, so long as she is truthful and just toward some people. She can participate in some social unions, while not really taking part in the social union of social unions. If her loves are not all-encompassing, she can protect those she cares about, or divert resources to them, while being unjust to people about whom she cares much less. And in a large society like the WOS, we would expect that each person's loves would *not* be all-encompassing or that concern for others would drop off with social distance. That, as we saw in §VI.6, is why Rawls cannot count on each person's balance of reasons being tipped toward justice by friendship, and why he cannot count on the *mutual assurance problem* being as easily solved as it would be were the WOS much smaller. Thus even if Joan wants to avoid deception, protect those close to her, and participate in social unions—even if C_4b, C_4c, and C_4d are true—satisfying those desires is compatible with her *not* treating her desire to act from principles of justice as supremely regulative. Rawls presented the *Argument from Love and Justice* as if this were not so "for simplicity." Once the simplifying assumption is dropped, the argument has an obvious weakness.

The weakness of the *Argument from Love and Justice* stems from the desires to which it appeals. Hypocrisy and deception seem to be moral failures. Friendship seems to be a moral relation. The desires to avoid hypocrisy and deception, and to live as friends, might therefore seem to be desires that move us to act as—or "to express our nature as"—moral persons. If they were desires that moved us to act as moral persons, then they would move us to be

just. But even if these desires do move us act from principles of justice sometimes or in our conduct toward some people, they do not reliably move us to be just persons—to be persons who always act from principles of right and who preserve our desire to regulate the whole of our lives by such principles. It follows that the desires to avoid hypocrisy and deception and engage in friendship do not, as such, move us to live as moral persons after all. They can move us to live as such persons if they are themselves parts of plan that is governed by higher-order desires to, for example, conduct our friendships in some ways and not others. But the reliance of the *Argument from Love and Justice* on desires that do not, as such, move us to live as moral persons is the source of its weakness.

To remedy the weakness, Rawls appeals to a "*special feature* of our desire to express our nature as moral persons"—the desire referred to by C_4a, which says that members of the WOS want to express their nature as free and equal rational persons. In the passage I quoted at the beginning of the last section, Rawls says that that desire "can be fulfilled only by acting on the principles of right and justice as having first priority." So what was assumed for simplicity's sake to be true of the desires appealed to by the *Argument from Love and Justice* really is true of the desire referred to by the *Kantian Congruence Argument*. The conclusion C_N can be "strengthen[ed]" if we follow the line of thought plotted by the *Argument from Love and Justice*, but substitute a premise about the desire to express our nature as free, equal, and rational for premises about the desires referred to by C_4b, C_4c, and C_4d.

This substitution enables the *Kantian Congruence Argument* to "strengthen" the conclusion of the *Argument from Love and Justice* C_N by placing it on a firmer footing than the *Argument from Love and Justice* did. But I think Rawls's appeal to "one special feature of the desire to express our nature" is supposed to strengthen C_N in another way as well. To see how else the appeal is supposed to strengthen it, we need to look again at just what the desire to express our nature provides a reason to do. That requires us to return to the argument from C_4a that I laid out in §V.3.

That argument begins, of course, with C_4a—a claim that, as we saw in §V.2, Rawls justifies by appealing to the Aristotelian Principle. That claim says:

> C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

The second and third steps of the argument are Rawls's assumptions that:

> (5.2) The desire to express our nature is a desire to act on principles that would be chosen in the OP.

and that

> (5.3) The desire to act justly is the desire to act on the principles that would be chosen in the OP.

(5.2) and (5.3) imply:

(5.4) The desire to express our nature has the same object as the desire to act justly.

I have treated the argument as concerned with a typical member Joan of the WOS. Applied to Joan, (5.4) implies:

(5.5) Joan can satisfy the desire asserted in C_4a by and only by acting justly.

Because the publicity condition is satisfied in the WOS and because moral learning in the WOS is transparent,

(5.6) "we are entitled to assume that [the] members [of the WOS] have a lucid grasp of the public conception of justice upon which their relations are founded" (*TJ*, p. 572/501).

So Joan, like everyone in the WOS, knows that (5.5) is true. Since she knows that her desire to express her nature is a desire to act from principles of justice:

(5.7) Joan's desire to express her nature moves her to act justly.

So:

(5.8) "when someone has true beliefs and a correct understanding of the theory of justice, these two desires move him in the same way" (*TJ*, p. 572/501).

And from this it follows that:

(5.9) "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*, p. 572/501).

This is the conclusion reached at the end of §V.3. When I looked at the argument carefully at the end of Chapter V, I pointed out that (5.9) seems to be too weak to support the conclusions Rawls wants to reach. Establishing (5.9) shows that insofar as members of the WOS want to express their nature, they have reason to act justly. But what Rawls wants to show is that their desire to express their nature gives them reason—ultimately, decisive reason—to treat their sense of justice as regulative and to do what they need to do to preserve that sentiment. As it stands, the argument does not seem to be nearly strong enough for that.

In §V.5, I suggested that Rawls could reach the stronger conclusions that he wants if, instead of appealing to (5.5), he could appeal to:

(5.5') Joan can satisfy the desire asserted in C_4a by and only by treating her sense of justice as supremely regulative of her other desires.

If Rawls could establish (5.5'), it—together with Rawls's assumptions about the publicity condition and its effects—would enable him to establish

(5.7') Joan's desire to express her nature moves her to treat her sense of justice as supremely regulative of her other desires.

(5.7'), together with (5.8), implies:

(5.9') "The desire to [treat our sense of justice as supremely regulative of our other desires] and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire" (*TJ*, p. 572/501).

In §V.2, I argued that the desire to express our nature as free and equal moral persons is not a desire we satisfy once and for all, nor is it a desire to show what we are at especially important moments in our lives. Rather, it is a desire that we try to satisfy over the course of life by deliberating and acting as free and equal rational beings. If that argument was right then, taken together with (5.9'), it implies that the desire to treat our sense of justice as supremely regulative is also a desire that we try to satisfy over the course of life, in this case by deliberating and acting as just persons. To deliberate and act as just persons over the course of life, we have to have an enduring sense of justice. A regulative sense of justice is a quality of character that can endure only with our work and commitment. It is therefore a sentiment that we must plan to preserve if we want to satisfy the desire to express our nature. This conclusion is what I have referred to as "the ostensible conclusion" of the *Kantian Congruence Argument*:

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 paragraph shows how Rawls gets to this conclusion from (5.9').

The defense of (5.5') is, I believe, the most philosophically interesting and ambitious part of the *Kantian Congruence Argument*. In §VII.4, I shall look at how Rawls defends it. But first note that even if Rawls can establish (5.5') and get to the ostensible conclusion, he would only have established that Joan knows she has a reason to maintain her sense of justice as supremely regulative. He would not have established congruence, for he would not have shown that we have *decisive* reason to express our nature. That crucial part of the *Kantian Congruence Argument* is assumed, rather than explicitly provided, in the text of the argument quoted above. The question is how Rawls fills it in. I shall address that question in the next section.

§VII.3: From the Ostensible Conclusion to Congruence

On one possible reading of the *Kantian Congruence Argument*, it is a very short step from the ostensible conclusion to congruence. The ostensible conclusion expresses a conditional which says that planning to preserve a regulative sense of justice is necessary if we are to realize our nature. While the truth-conditions

of the antecedent of the conditional are not obvious, it would be natural to interpret the conditional as saying what we must do if we *want* to express our nature. According to C_4a, members of the WOS *do* want to express their nature. Conjoined with the ostensible conclusion, C_4a therefore implies that members of the WOS would in fact maintain their sense of justice as supremely regulative. Since the sense of justice is a desire to act on principles of justice, then—provided that the decision to maintain the sense of justice is reached within the appropriate viewpoint—Rawls could infer the conclusion that I have said some readers think he reached:

> C_C': Each member of the WOS judges, from the viewpoint of full deliberative rationality, that she should treat her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

This may seem an initially plausible reading of the *Kantian Congruence Argument*. Its prima facie plausibility draws further support from the fact that, as I have already said in sketching my own reading, the *Kantian Congruence Argument* opens with a promise to strengthen the conclusion of the *Argument from Love and Justice*. The alternative reading I am now considering shows that that promise can be fulfilled in an elegant and economical way. For in Chapter VI, I contended that that argument supports congruence by way of *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

This, I said, is the conclusion to be strengthened. C_C' is a stronger claim than C_N, since it does not include a qualifier saying that members of the WOS would maintain the desire to act from the principles as supremely regulative when others do so as well. On the reading of the *Kantian Congruence Argument* I am now considering, the argument moves directly from the ostensible conclusion to congruence, while bypassing *TJ's Nash Claim*. The promise to "strengthen" the conclusion of the *Argument from Love and Justice* is thus to be read as the promise to reach the stronger conclusion C_C' without relying on the weaker C_N.

This reading may be attractive because it does not require us to interpolate much argument between the ostensible conclusion of the *Kantian Congruence Argument* and the conclusion that the right and the good are congruent. Moreover, showing congruence is supposed to show that members of the WOS have an effective "desire to conduct [themselves] in a certain way above all else" (*TJ*, p. 574/503)—an effective desire to be the kind of person who refuses to trade off considerations of justice against other concerns when she adopts the viewpoint of full deliberative rationality. On this interpretation, members of the

WOS do not consider their balances of reasons even in deciding whether to be that kind of person. They want to express their nature "above all else." Having "a lucid grasp of the public conception of justice" (*TJ*, p. 572/501), they realize how they must live if they are to express it. Considerations that tell in favor of being a different kind of person are, as it were, evacuated of their force.

There are, however, a number of difficulties with this interpretation.

First, the impression that, on this interpretation, members of the WOS decide to maintain their sense of justice without balancing their reasons is an illusion, since they make their decision because of the strength of their desire to express their nature. Without some such assumption, the reason provided by that desire cannot be shown to be decisive. But the assumption that that reason is strong just is the assumption that it is weighty enough to tip the balance against countervailing considerations. On the reading now under consideration, it is not at all clear how that assumption could be defended.

Moreover, someone who adopts the viewpoint of full deliberative rationality takes all her desires into account, including her effective desire to be the kind of person who does not balance considerations of justice against other considerations. It is true that there would be some incongruity in deciding to be that kind of person on the basis of one's balance of reasons, *if the decision were made from the viewpoint of full deliberative rationality*. The avoidance of this incongruity may seem to be one of the advantages of the interpretation I am now considering. But the congruence arguments concern someone who follows the thin theory of the good, not someone judging from the viewpoint of full deliberative rationality. The person who follows the thin theory leaves out of account her desire to be just under that description. There is no more incongruity in her deciding from the point of view of the thin theory that her balance of reasons tells in favor of being a just person than there would be in deciding from a self-interested point of view that being just is to her advantage narrowly construed. What looked like an advantage of the alternative interpretation ceases to seem like one once we recall the viewpoint from which the relevant decision is made.

Finally, if this reading of the *Kantian Congruence Argument* is right, then the argument is open to a very serious objection. To see this, recall that according to the ostensible conclusion of that argument, "in order to realize our nature we have no alternative but to plan to preserve our sense of justice as governing our other aims." That means that the person who wants to express her nature will take her sense of justice as supremely regulative *regardless of what she thinks others will do*. To trade off her sense of justice, even when others are unjust, would be "to give way to the contingencies and accidents of the world" (*TJ*, p. 575/503). The interpretation I am now considering moves from this claim to C_C', via the assumption that the sense of justice is a desire to act from principles of justice. If the phrase "the principles of justice" in C_C' is understood, as it must be, to refer to the principles of justice *for a WOS*, then that is how the phrase must be understood in the assumption too. And so on this reading of the argument, what the ostensible conclusion shows is that if

members of the WOS want to express their nature, they will take the desire to act from *those principles* as supremely regulative, regardless of what they think others will do. To put the point in game-theoretic terms: taking the desire to act from those principles as supremely regulative is their dominant strategy. It is precisely by taking the ostensible conclusion to show this that the reading I am now looking at bypasses *TJ's Nash Claim* and the *mutual congruence problem*, and move directly to congruence.

As I noted in §II.3, the claim that someone would think it rational to regulate her plans by the principles of justice for a WOS regardless of what she thinks others will do is very strong, since doing so would leave her liable to very serious losses in event that others are unjust. Since the *Kantian Congruence Argument* concerns someone "following the thin the theory," the payoff for preserving the desire to act from principles as supremely regulative must be goods that are good according to the thin theory. It is far from clear how those goods, even the good of living as a free and equal rational person, could be valued so highly. And it is doubtful that Rawls thinks they can be, since he says that "even with a sense of justice"—which I take to imply "even with desires for far more than can be reckoned good according to the thin theory"—"men's compliance with a cooperative venture is predicated on the belief that others will do their part" (*TJ*, p. 336/296).

Thus the interpretation of the *Kantian Congruence Argument* I have been considering fails to show why members of the WOS take their desire to express their nature as decisive, and it implausibly says that members of the WOS would regulate their plans by the principles of justice chosen for a WOS, regardless of what they think others will do. I now want to fill in some details of the interpretation I sketched in §VII.1, showing how—on my reading—Rawls moves from the ostensible conclusion to congruence while avoiding these difficulties.

In Chapter V, we saw that even when Joan, the typical member of the WOS, follows the thin theory of the good, she still has several reasons to be a just person. Some of those reasons stem from her desires for the goods of friendship and association. Because the values of these goods are accounted for by the thin theory, the reasons that stem from them are what I called "thin reasons to be just." In Chapter VI, we saw how Rawls tries to show that those reasons are decisive. He does not—as on the interpretation I just considered—simply assume that Joan's desires for these goods are strong enough to tip her balance of reasons. He relies on a *Balance Conditional*. The argument for congruence then went by way of an argument that Joan would not regret maintaining her sense of justice.

On my reading, the *Kantian Congruence Argument* follows the pattern laid down there. In the last section, we saw how Rawls gets to:

> (5.9') "The desire to [treat our sense of justice as supremely regulative of our other desires] and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire."

We also saw that it is a short step from (5.9') to the "ostensible conclusion" of the *Kantian Congruence Argument*:

> 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.

How does Rawls show that the desire to express our nature is decisive, so that he can infer the congruence of the right and the good?

I take Rawls to argue that even in the world as it is, with all its injustice, Joan would rather live as a free and equal rational being than not, and that failing to live as such a being is something she would deeply regret. According to the ostensible conclusion, she can live as a free and equal rational being only if she preserves her sense of justice as supremely regulative. So even in the world as it is, the balance of her thin reasons tips in favor of preserving her sense of justice. Rawls then argues that if her desire to live as a free and equal rational being tips her balance toward justice in the world as it is, it tips her balance toward being just in a WOS. To sum up this line of reasoning: Rawls moves from the ostensible conclusion to the claim that Joan would in fact maintain her sense of justice via a *Balance Conditional* that says:

> If Joan's balance of reasons would tilt in favor of preserving her sense of justice as supremely regulative in the world as it is, then her balance of reasons tilts in favor of preserving it as supremely regulative in the WOS.

To see why Rawls thinks the antecedent of this *Balance Conditional* is true, note first that Joan would have a reason to express her nature in the world as it is—a reason stemming from C_4a. For when I argued for C_4a in §IV.2, I noted that while the development of a *free-and-equal self-conception* and the emergence of a desire to express our nature may depend upon a liberal democratic culture, they do not depend upon the distinctive conditions of a WOS. If Joan lived in liberal democratic societies as they are, she would have a *free-and-equal self-conception* and would want to live as a free and equal person—as I believe Rawls thinks those of us who are citizens of such societies actually do. We, the Rawls of *TJ* believed, think of ourselves as free and equal rational beings and want to live that way. We, he thought, have a desire to express our nature. So in the world as it is, Joan would have a reason to express her nature too.

Moreover, I believe Rawls thinks that in the world as it is, we would rather express our nature than not, and express it by doing what (5.5') and the ostensible conclusion imply that we must do if we are going to act as free and equal rational beings: govern ourselves by our sense of right. Our recognition of this may not be explicit but it is, Rawls says, testified to by our moral sentiments. For in laying out the *Kantian Congruence Argument*, Rawls says that "acting wrongly is *always* liable to arouse feelings of guilt and shame." I take the force of the "always" to be that acting wrongly renders us liable to feelings of guilt and shame in the world as it is, and not just in the WOS. Rawls is quite clear that shame is the natural response to recognition that we have not expressed our nature, but have "acted as though we belonged

to a lower order" (*TJ*, p. 256/225). It would be inappropriate to feel shame when we act against the principles if expressing our nature were not something we had decisive reason to do. And so the propriety of the feeling of shame when we act wrongly in the world as it is shows, I believe, that we—and hence Joan—would rather express our nature even in the world as it is than be the kind of people who decide whether or not be just as suits our convenience. From this, and the ostensible conclusion, the antecedent of the *Balance Conditional* follows.

Why does it follow that Joan's balance of reasons would tilt in favor of preserving her sense of justice as supremely regulative in the WOS?

My argument for the antecedent of the relevant *Balance Conditional* depended upon the assumptions that we who live in liberal democratic societies in the world as it is have at least an implicit grasp of our nature as free and equal persons and that we recognize that we fail to live up to that conception of ourselves when we act unjustly. But as (5.6) says, members of the WOS have a "lucid grasp" of the public conception of justice. I assume this implies that they have, if anything, a clearer understanding of their nature as free and equal persons than many of us do in the world as it is, and a clearer understanding of the connection between expressing their nature and acting from the principles of justice. If we in the world as it is recognize that violating the principles of justice is a failure to live up to what we can be, members of the WOS must recognize this even more clearly.

And so if Joan would know (5.5') and the ostensible conclusion in the WOS, then she would know that she must treat her sense of justice as supremely regulative if she is to express her nature. She would also know that the sense of justice, as Rawls says, "reveals what a person is" and that to compromise it by treating it "as but one desire to be weighed against others is not to achieve for the self free reign but to give way to the contingencies and accidents of the world" (*TJ*, p. 575/503). Joan would therefore know that if she did trade off her desire to be just in this way, her life would betray—rather than express—what she is. This knowledge would affect what Joan would regret doing. Knowing (5.5') and the ostensible conclusion, she would know that what she would regret is not a life in which she maintains her sense of justice as supremely regulative, but one in which she stands ready to compromise it. So the worry Joan would have about a commitment to being a just person is allayed. Being just is something she—and others—have reason to do, and it is something they would not regret doing if they made that commitment. This, Rawls thinks, is enough to show that in the WOS, Joan's reasons for expressing her nature tell decisively in favor of maintaining her sense of justice.

From this, and the fact that Joan is typical, Rawls can infer:

(5.12) Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans.

The *Balance Conditional* enables Rawls to infer (5.12) without assuming—as on the alternative reading that I considered—that Joan simply attaches very great weight to the desire to express her nature.

Intuitively put, (5.12) says that even from within the thin theory, Joan would judge that it is good to be just, regardless of what she think others will do. But (5.12) does not imply that in order to realize her nature, Joan would treat *her desire to act from principles of ideal theory* as supremely regulative, regardless of what she thinks others will do. What it implies is that she would treat *her sense of justice* as supremely regulative, regardless of what she thinks others will do. Nothing Rawls says in *TJ* prevents him from saying that parties in the OP choose one set of principles as part of ideal theory, and different and less-demanding principles of right as part of nonideal theory.

Rawls does not explore Kantian arguments for non-ideal principles, but Christine Korsgaard does and draws on some of Rawls's conceptual apparatus to do so.[7] I shall not examine her discussion here; I shall, however, assume that Rawls would allow for nonideal principles on something like the grounds Korsgaard provides. Now suppose that Joan's reasons to express her nature always outweigh countervailing considerations. Then she will judge from within the thin theory that she should regulate her life by her sense of justice regardless of what others do, as (5.12) says. But because she can express her nature by acting from different principles in ideal and non-ideal circumstances, Rawls is not committed to the implausible view that she will judge it rational to express her nature by acting from ideal principles come what may. Thus, my reading avoids the second of the difficulties that afflicts the interpretation I considered at the beginning of this section.

The price of avoiding that difficulty is that my reading opens the argumentative distance to which I referred earlier between the ostensible conclusion and (5.12) on the one hand, and congruence on the other. It does so because the principles of ideal theory are the principles of justice that would be chosen in the OP to regulate the basic structure of a WOS. So while a person who wants to express her nature might not regulate her life by those principles under any circumstances whatever, she would regulate her life by them if she knew she was in the special circumstances of the WOS. The conclusion of the congruence arguments is supposed to imply that she would maintain her desire to act from those principles, the principles of ideal theory. Thus if my reading of the *Kantian Congruence Argument* is right, then even after Rawls establishes (5.12) he must still be concerned to show that Joan knows that her society is well-ordered. On my reading of the argument, he is.

To see this, note first that it is a short step from (5.12) to *TJ's Nash Claim*. For (5.12) implies that Joan would decide to maintain her desire to act from principles of justice which are appropriate to the circumstances in which she

7. Christine Korsgaard, "The Right to Lie: Kant on Dealing with Evil," *Philosophy and Public Affairs* 15 (1986): pp. 325–49.

thinks she finds herself. Since Rawls's two principles are the principles of justice for a WOS, someone with a sense of justice would—as I have said—decide to maintain her desire to act from those principles if she believed her society to be well-ordered. A WOS is a society in which each person knows that everyone else regulates her life by the principles of ideal theory (see *TJ*, p. 5/4). So (5.12), together with the definition of a WOS, implies that Joan would maintain her desire to regulate her life by the principles of ideal rather than nonideal theory when she knows that others are committed to regulating their lives by those principles as well. This fact about Joan, together with the fact that Joan is typical, implies *TJ's Nash Claim*, where the phrase "the principles" refers to Rawls's principles of justice:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

To show stability, Rawls needs to reach:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

And he wants to move from C_6 to the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from within the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

When we looked at the *Argument from Love and Justice* in Chapter VI, we saw that to reach his desired conclusions, Rawls still had to address the *mutual assurance problem* even after he established *TJ's Nash Claim*. The same is true of the *Kantian Congruence Argument*. To move from C_N to his desired conclusions, Rawls needed to show how each member of the WOS can be assured that others—or almost all others—value expressing their nature highly enough to affirm their sense of justice, so that each is assured that the conditions of ideal theory obtain and will continue to obtain.

We saw at the end of Chapter VI that there is some difficulty in supposing that the *mutual assurance problem* is solved by each person's perceiving the "evident intention" of everyone else in the WOS to promote the good of others. The WOS is too large for that, and its size raises the question of how each member of the WOS could know that others valued friendship enough to treat others justly. Once we see how the *Kantian Congruence Argument* goes, however, the solution to the *mutual assurance problem*—as raised by that argument—is supposed to be clear.

The second step in the *Kantian Congruence Argument* says that:

(5.2) The desire to express our nature is a desire to act on principles that would be chosen in the OP.

And so to see that others value expressing their nature, it is sufficient to see that they in fact live justly. This is something each person in the WOS is in a position to see. Each person can infer a good deal about the just conduct of others by seeing the low incidence of crime and cheating; this inference provides each person some assurance that she is in ideal circumstances. Moreover, each can see that conflicting claims are adjudicated from a "unified perspective" provided by "common allegiance to justice" (*TJ*, p. 474/415)—by which I presume Rawls means that in the WOS, everyone knows that competing claims are settled by appeal to a conception of justice that is commonly accepted as giving the final disposition of the matter. Each can see, then, that her society is regulated by justice as fairness. And each can see that it conforms to justice as fairness with minimal reliance on coercion (*TJ*, pp. 575–76/504). This is possible only if a sufficiently large number of people supported those institutions from a sense of justice. Thus, common knowledge that the WOS is just shows each that everyone else values living as a free and equal person, and solves the *mutual assurance problem.* Moreover, since acting justly when others do makes great goods available and since a sense of justice can only uprooted with some difficulty, each person would rather maintain it than not. All each person needs to maintain it is the assurance that others have a sense of justice that they prefer to maintain. And once the *mutual assurance problem* is solved, Rawls can move from *TJ's Nash Claim* to C_6. We have already seen why he can move from C_6 to the *Congruence Conclusion.* This move completes the *Kantian Congruence Argument.*

I believe Rawls thinks that the part of the *Kantian Congruence Argument* that I have surveyed in this section could go largely without saying. The move from (5.9) and the ostensible conclusion to (5.12) requires appeal to a *Balance Conditional*, but Rawls may have thought his readers could supply it for themselves once they had seen a similar appeal in the *Argument from Love and Justice*. The argument from (5.12) to the *Congruence Conclusion* just required bearing in mind the distinction between ideal and non-ideal theory and the special circumstances of a WOS, but Rawls could have expected his readers to do that.

The really interesting part of the *Kantian Congruence Argument* is the part actually laid out in the text, which moves from C_4a to the ostensible conclusion. In the next section, I will look at how Rawls supports the critical steps in that part of the argument. Before doing so, however, I want to note a couple of other points about the part of the argument I have reconstructed so far.

Recall that the *Kantian Congruence Argument* opened with the promise to strengthen the conclusion of the *Argument from Love and Justice*. On my reading, that is not simply the promise to put that conclusion on a firmer footing by appealing to C_4a and the desire to express our nature. But neither is it—as on the alternative reading I considered earlier—the promise to bypass

TJ's Nash Claim and the *mutual assurance problem*. It is the promise to support the conclusion of the *Argument from Love and Justice* by showing that it follows from a strong but plausible claim about what we must do to express our nature. That claim is the ostensible conclusion, which says that to express our nature, members of the WOS must preserve their sense of justice as supremely regulative, regardless of what others do.

In arguing for the consequent of the *Balance Conditional* by which the *Kantian Congruence Argument* moves to (5.12), I assumed that Joan has some understanding of her own nature as a free and equal rational person, and of the connections between expressing her nature and treating her sense of justice as supremely regulative. I assumed, that is, that she has and wants to live up to a certain view of herself—what Korsgaard calls a "practical identity" and what I called in §IV.2 a *self-conception*. We know by C_4a that Joan has, and wants to live up to, what I called the *free-and-equal self-conception.* She thinks of herself, and wants to act as, what (1.1) of the Pivotal Argument says she is—a free and equal rational being capable of reflecting on the ends she pursues. But the self-conception presupposed by the *Kantian Congruence Argument* must be more demanding than that. For the *Kantian Congruence Argument* can succeed only if Joan is liable to shame if she "give[s] way to the contingencies and accidents of the world." She is liable to shame for doing *that* only if she wants to live up to *one particular conception* of her freedom. And so she must know that if she wants to express her nature, then she will have to live as a being who is free in that way. The question is what the relevant conception of freedom is.

Joan's desire to live up to the conception of freedom at work in the *Kantian Congruence Argument* is a conception-dependent desire. It is important, however, that it is not an "ideal-dependent desire" as I have used that term. In particular, it is important that it is not a desire to live as a fully autonomous person. For the values of the objects of ideal-dependent desires—such as the desire to be fully autonomous—are given by the full theory of the good, and so presuppose the content of the principles of justice. I have supposed that Joan has such desires and that they may govern her action in daily life. But the *Kantian Congruence Argument*, because it is intended to answer the problem of congruence in its non-trivial form, presupposes that Joan adopts a certain perspective on her desires and on herself. For purposes of argument, she follows the thin theory of the good rather than the full theory. The *Kantian Congruence Argument* cannot, therefore, depend upon Joan's desire to be fully autonomous. So while Joan must value a particular conception of her freedom if the *Kantian Congruence Argument* is to succeed, the freedom that she is assumed to value for purposes of that argument is not full autonomy or any other conception of freedom that depends upon the content of the principles.

The *Kantian Congruence Argument* is to establish that Joan's concern with this kind of freedom gives her reason to act from principles chosen in the OP. So the conception of freedom that Joan is presumed to value for purposes of that argument must be the freedom realized when one acts on principles chosen subject to the conditions that make choice in the OP free choice. The

content of Rawls's principles contributes to the freedom someone realizes when she acts on them. But because Joan is following the thin theory, the value she attaches to acting from the principles cannot depend upon that contribution. Insofar as she values the freedom she realizes by acting on the principles, what she must value is the freedom realized by acting from "first principles [that] are [*not*] decided by natural contingencies" (*TJ*, p. 256/225). She must value the freedom she realizes by acting from principles chosen subject to the conditions of the OP. That is the kind of freedom that I called *thin autonomy* in §III.2.

If the *Kantian Congruence Argument* is to succeed, Joan must have this conception of her freedom. She must value it when she adopts the perspective on her desires and herself that the congruence arguments require. She must be in a position to know that the best way for someone following the thin theory to satisfy the desire referred to by C_4a—and to live up to the *free-and-equal self-conception*—is to realize the associated kind of freedom in her action. And she must be in a position to know that the best way for her to realize that kind of freedom in action is to act while taking the principles as supremely regulative. But, it might be said, if what Joan is presumed to care about for purposes of the *Kantian Congruence Argument* is acting from principles that would be chosen in the OP regardless of their content, how does the argument establish that Rawls's two principles—as opposed to some others that might be adopted—are congruent with the good?

In reply, Rawls would point out that the first part of *TJ* shows that the two principles would be adopted in the OP. The treatment of congruence assumes that that is already established (*TJ*, p. 567/497). The nontrivial question of congruence asks, in effect, why those principles are congruent with the good apart from any desire to act from principles with their distinctive content. An argument for congruence that does not appeal to the good of acting from principles with their content is just what we would expect, given the way the treatment of congruence is set up. Moreover, it should not be surprising that a contract view would locate some of the value of acting from principles in the way that those principles were chosen. That is just what the *Kantian Congruence Argument* does, on my reading.

When I laid out the *Kantian Congruence Argument*, I said that Joan has a clear understanding of her nature because she lives in the WOS, where the publicity condition is satisfied and moral education is transparent. This suggests that the publicity of the public conception of justice, and its educative role, are responsible for Joan's coming to understand the conception of freedom at work in the argument. The argument as a whole suggests that what Joan understands about her nature affects how she wants to express her nature.

This latter point, too, is just what we should expect, given the way Rawls would argue that we want to express our nature. That argument begins from the *Two Conjunct Reading* of the Aristotelian Principle. According to that interpretation of the Principle, the desire to exercise our natural powers is sensitive to what we come to believe about what we are and about how it is

natural for us to act. We shall see later that Rawls's treatment of publicity and the educative role of justice in the original *Dewey Lectures* confirm this suggestion. We shall also see why Rawls made the changes between *TJ* and *PL* as a result of thinking more deeply about how publicity helps us to grasp our own nature. To understand this, we have to see why Rawls makes the crucial move in the *Kantian Congruence Argument* as I have reconstructed it, (5.5').

§VII.4 Establishing (5.5')

Recall that (5.5') says:

(5.5') Joan can satisfy the desire asserted in C_4a by and only by treating her sense of justice as supremely regulative of her other desires.

How might an argument for this claim go? Because the sense of justice is the desire to act from principles that would be chosen in the OP, Rawls could establish (5.5') if he could show that:

(5.2.1) We can satisfy the desire to act from principles chosen in the OP only if we treat that desire as supremely regulative.

For according to (5.2):

(5.2) The desire to express our nature is a desire to act from principles that would be chosen in the OP.

So (5.2.1)—together with (5.2)—would enable Rawls to infer:

(5.2.2) The desire to express our nature can be satisfied only if we treat the desire to act from principles that would be chosen in the OP as supremely regulative.

(5.2.2) does not itself imply (5.5'), for (5.2.2) states only a necessary condition of satisfying the desire to express our nature. But (5.4) says that the desire to express our nature has the same object as the desire to be just, so we know that we can satisfy the one desire by and only by satisfying the other. And we know that we can satisfy the desire to be just by acting from principles that would be chosen in the OP. What (5.2.1) adds to this conclusion is that we can satisfy the desire to be just only if we do not merely act from the principles but treat the desire to act from them as supremely regulative. So we know that we can satisfy the desire to express our nature by and only by satisfying the desire to act justly, and that the only—and hence the best—way to satisfy *that* desire is to treat the desire to act from the principles, and hence the sense of justice, as supremely regulative. This gets us to (5.5').

Can Rawls establish the claims he needs to infer (5.5') and to vindicate the *Kantian Congruence Argument*? He seems quite explicitly to endorse the equivalent of (5.2.2) in the course of laying out the *Argument*, for he says:

"The desire to express our nature as a free and equal rational being can be fulfilled only by acting on the principles of justice as having first priority" (*TJ*, p. 574/503). He then adds immediately:

> This is a consequence of the condition of finality: since these principles are regulative, the desire to act on them is satisfied only to the extent that it is likewise regulative with respect to other desires.

I take the phrase "the desire to act on them is satisfied only to the extent that it is likewise regulative with respect to other desires" to imply (5.2.1). So this remark confirms that Rawls's defense of (5.2.2) does indeed go by way, if not exactly of (5.2.1), then of a thesis that implies it.

The quoted remark also indicates that (5.2.2) ultimately depends upon the finality condition. That condition is a condition on principles adopted in the OP. The appeal to finality here should not, I think, be read as asserting an additional premise. Rather, it should be read as reminding us of just what premise (5.2) really says. I phrased (5.2) as I did—as referring to "principles that would be chosen in the OP"—in deference to the way Rawls puts things when he lays out Joan's reasons to be just. But what he *means* by (5.2) is that the desire to express our nature is a desire to act on principles chosen *subject to the conditions of the OP*—including the finality condition. If someone acknowledges principles as final she will, Rawls thinks, she treats their desire to comply with the principles as "regulative with respect to other desires" (*TJ*, p. 574/503). So the just person must treat those principles as regulating his deliberation about what ends to pursue and how to pursue them.[8] And so

8. Why is this so?

A just person living under just institutions wants to be the sort of person who accepts institutional verdicts because they are just, so he must want to be the sort of person who does not want or plan to appeal the verdicts in ways that are unreasonable. If his sense of justice is effective, it limits his plans and desires at least to that extent. Furthermore, if he really does want to recognize the verdicts of institutions as just, then he must want to base his plans and his claims on the verdicts institutions have rendered in his own and other cases. He must also want the claims he advances to be based on the reasonable expectations and desires he has formed, based on the principles of justice and on his knowledge of how institutions have complied with them. He must therefore want to be the kind of person who recognizes a distinction between what he can claim from institutions and his fellow citizens and what he merely wants those institutions and his fellow citizens to do. And he must want to be the sort of person who does not wish to advance claims to advantages simply because he desires the advantages he would enjoy if those claims were honored.

Grant that this kind of self-discipline—the self-discipline exercised by someone who does not move immediately to claims from desires, however intense—can appropriately be described as a "regulation" of desire. Then the just person can satisfy his desire to act on principles which are final only if his desire to act from the principles regulates his other desires. This is why Rawls says at *TJ*, p. 574/503, that finality implies that the just person must treat the principles of justice as regulative.

(5.2), understood in the expansive way that I have just suggested, implies (5.2.1). So Rawls can infer (5.2.2) and hence (5.5').

There are many puzzling features of the *Kantian Congruence Argument*. Perhaps none is more vexing than this appeal to finality. The finality condition was originally introduced as one among several "formal constraints on the concept of right" (see *TJ*, §23). Nothing Rawls says when he introduces these constraints indicates that they are particularly important. The constraints are, Rawls says innocuously, "natural enough" (*TJ*, p. 131/113). The finality condition figures in Rawls's arguments for the principles in part I of *TJ*. (*TJ*, pp. 176ff./153ff.) After that, only scattered remarks in the intervening pages (e.g., *TJ*, p. 478/418) hint at the role finality is to assume in the treatment of congruence. Yet it turns out that this condition is ultimately supposed to support (5.5'), one of the critical claims in a very important argument. But the appeal to finality does not engender puzzlement simply because the importance of finality in the *Kantian Congruence Argument* is surprising. It also engenders puzzlement because of what the finality condition turns out to be important *for*.

I shall state the requirement of finality more fully at the beginning of §VII.6. For now, suffice it to say the finality condition requires that parties to the OP are to evaluate principles knowing that the principles they adopt will be the final arbiters of conflicting claims. If the principles of justice imply some solution to a question of justice, that solution is dispositive. There is no appeal to further principles to "check" the solution. Since it is natural enough to suppose that this is just the role principles of right play in our lives, finality may have seemed a natural enough condition to include in the OP when Rawls was using the OP to identify such principles.

But the *Kantian Congruence Argument* appeals to finality at a critical juncture to show that the principles chosen in the OP are such that taking them as supremely regulative belongs to Joan's *good*. Readers who granted the naturalness of finality and the other conditions on the OP when Rawls proposed using a social contract for one purpose may well be puzzled by his attempt to exploit one of those conditions for what seems to be a very different purpose altogether. After all, Rawls never invited us to consider finality or any other of the conditions of the OP with this purpose in mind. On the contrary, in introducing the formal constraints on the concept of right, he said "the propriety of these formal conditions is derived from the task of principles of right in adjusting the claims that persons make on their institutions and on one another" (*TJ*, p. 131/113).

Before we conclude that second thoughts about finality and the other conditions are in order, I want to look at Rawls's reasons for accepting (5.2). For whatever Rawls may have said when he introduced various conditions of the OP, we can see in retrospect that he framed the OP with an eye toward showing congruence by relying on (5.2). We can see, that is, that the conditions of the OP were chosen precisely to make it the appropriate device for identifying principles of right *and* to make the OP such that acting from

principles chosen there would be expressive of our nature and part of our good. In sum, they were chosen to enable the OP to play what I shall refer to as the *bridge function*, bridging the right and the good. That the OP plays this function bears on the question of whether the OP could indeed be eliminated from the arguments of *TJ*, a much-contested question to which I shall return at the end of this chapter.

We saw in §IV.2 that the desire to express our nature is a higher-order desire to form, revise, and execute our plans in ways that befit persons who are free, equal, and rational. Why should we think that we can satisfy that desire only if we satisfy the desire to comply with the principles we would adopt in the OP? Why, that is, should we accept (5.2)?

§VII.5: Defending (5.2)

Rawls's argument for (5.2) turns, I believe, on a claim that he makes in the section of *TJ* devoted to the Kantian Interpretation: the claim that "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*, p. 253/222). Let's call this claim the *KI Claim*.

The *KI Claim* is very general. It ranges over all kinds of beings that are capable of choosing principles and acting on them. It is not immediately clear what claim it asserts of such beings or what would count as confirming the claim. In part this is because Rawls does not say what he means by the phrase "act on principles", which could mean "act according to," "act from" or something else. Moreover, Rawls does not offer much by way of argument for the *KI Claim*. Any attempt to extract an argument from the text is bound to be highly speculative.

Human persons are clearly the instance of the *KI Claim* with which Rawls is most concerned. We can make some headway interpreting and defending the claim by considering that case. We know that to express our nature as free and equal rational beings is to conduct ourselves as such beings. So, applied to human beings, the *KI Claim* says that to conduct ourselves as free, equal and rational is to act on principles that would be chosen if our nature as free, equal, and rational were the decisive determining element of the choice of those principles. Perhaps what Rawls has in mind as a defense of the *KI Claim* is something like the following.

If a being B acts on some principle P, then at the very least B's action must be permitted by P; otherwise it is hard to see how B would be acting *on* the principle at all. Moreover, since B is assumed capable of reflection and choice, B must be capable of determining whether its action is permitted by P and choosing accordingly. Thus if P is a principle prohibiting theft, then—if I am to be said to act on this principle—my action must, at minimum, be consistent with

it. And since I am capable of choice, I act on the principle only if I am capable of checking to see that P permits my action and am disposed to choose my action in accord with what I find. Now suppose that if I were asked to choose principles to assess the permissibility of my conduct, one of the principles I would choose is P and I would choose P because of the kind of being I am—a being who lives among other beings, who has material needs, and who meets those needs through a regime of property holdings. In that case, my nature is the "decisive determining element" in the choice of P. Then when I act on P in the sense of "act on" just specified, I conduct myself in a way that suits the kind of being I am. My action on P expresses my nature, so the *KI Claim* is true.

It is clear from the context of the *KI Claim* that, as we would expect, Rawls thinks our nature as free and equal rational beings is "the decisive determining element" of the choice in the OP. This is a familiar claim. It is step (1.9) in what I called in Chapter I the Pivotal Argument of the *Public Basis View*. We saw then that proponents of the *Public Basis View* think Rawls relies on (1.9) to identify principles of justice. I said then that while Rawls does indeed rely on (1.9) for that purpose, we would see that Rawls also relies on it in a different connection. The other connection in which he relies on it is now clear. Rawls relies on (1.9) to move from the *KI Claim* to (5.2), a critical step in the *Kantian Congruence Argument*. That claim says that:

(5.2) The desire to express our nature is a desire to act from principles that would be chosen in the OP.

But what is it about freedom, rationality, and equality—and about the conditions of the OP—that licenses (1.9) and the move from the *KI Claim* to (5.2)? Which conditions of the OP make it the case that choice there is determined by the kind of beings we are?

The most controversial—and seemingly the most interesting—condition of the OP is the veil of ignorance. It is tempting to seize on the veil as providing the sole answer to my question for that reason, and because of Rawls's own remarks about the veil of ignorance in the passages in which he himself seems to anticipate and answer the question (cf. *TJ*, p. 252/222). Moreover, I argued above that the *Kantian Congruence Argument* succeeds only if Joan thinks she expresses her nature as free by realizing what I called *thin autonomy*, the freedom she realizes by acting from principles that are chosen subject to the conditions of the OP. The veil of ignorance is the element of the OP that insures that principles chosen there are chosen freely. Finally, the *Kantian Congruence Argument* is explicitly premised on the "desire to express our nature as *moral* persons" (*TJ*, p. 574/503, emphasis added). In light of Rawls's later explanation of how the veil insures that parties to the OP are equally situated as such persons,[9] it would be natural to stress the veil in answer to my question for that reason as well.

9. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, pp. 316–17.

But it is also tempting to seize on the veil as the answer because it is tempting to read "rational" out of the question and take it to be asking what the conditions of the OP are in virtue of which we express our nature as free and equal beings by acting from principles chosen subject to those conditions. After all, we may think, the rationality of the parties is given by a set of stipulations that are quite straightforward (*TJ*, §25). Our nature as free and equal determines choice in the OP, in much more interesting and complicated ways, by the veil of ignorance. And so, we may conclude, the interesting answer to the question must lie there.

This answer misfires in two ways. It is true that a full answer to the question I have posed about the OP would treat of the connection between freedom and the veil of ignorance. This is a point to which I shall return in §VII.8. But to suppose that the veil of ignorance provides the *sole* answer overlooks the fact that other conditions on the OP that determine choice also represent elements of our nature.[10] Furthermore, the attempt to read "rational" out of the question rests on the assumption that our rational nature is represented in the OP simply by the way the parties compare and choose among the commodity bundles that would be available to them under various conceptions of justice. This assumption is mistaken. Our rational nature includes the ability to do far more than engage in the kind of reasoning the parties engage in, and our rational nature is represented by more conditions of the OP than the parties' powers of reasoning. My attempt to defend (1.9) will therefore presuppose that when Rawls speaks of our desire to express our nature as free and equal rational beings, he is not using "rational" in the technical sense in which it is opposed to the "reasonable." Rather, it refers also to the interests and powers we have as beings who exercise practical reason.

Unfortunately, I cannot give a full defense of (1.9) and the move from the *KI Claim* to (5.2) here. I am going to ask about the conditions of the OP in virtue of which we express our nature as *rational* by acting from principles chosen subject to those conditions. Doing so not only combats the tendency to read the critical adjective "rational" out of those claims, it also enables us to see how the *Kantian Congruence Argument* fits into its immediate context. Since we want to understand Rawls's appeal to finality in the *Kantian Congruence Argument*, I want to zero in on the connection between that condition and our rational nature.

10. See, for instance, Rawls's remark about the unanimity condition at *TJ*, p. 564/494: "the unanimity condition on principles of justice is suited to express even the nature of a single self." In later work, Rawls also emphasizes—in a way that *TJ* did not—the connection between our nature and the parties' desire for primary goods; see, for example, "Kantian Constructivism in Moral Theory," *Collected Papers*, Lecture 1.

§VII.6: Finality, Rationality, and the Unity of the Self

Rawls uses the term "finality" and its cognates to denote two different conditions on the choice of principles.[11] The first of these is a condition on the *principles to be chosen* in the OP and is spelled out in the section of *TJ* devoted to "The Formal Constraints on the Concept of Right." We can call it the condition of *ultimacy*. The principles chosen in the OP are to be "the final court of appeal in practical reasoning. There are no higher standards to which arguments in support of claims can be addressed; reasoning successfully from these principles is conclusive" (*TJ*, p. 135/116). So if two people put forward competing claims of one another or of their institutions, the principles chosen in the OP are to be the final or the ultimate standard by which those claims are to be settled. There can be no "checking" of the result by appeal to further ethical principles. The second is mentioned several sections later as a condition on the *choice*. Rawls says of the agreement reached in the OP that because it is "final and made in perpetuity, there is no second chance" (*TJ*, p. 176/153). The context strongly suggests that here, Rawls takes "final" to be synonymous with "made in perpetuity."[12] The condition says that parties in the OP are choosing the principles that will regulate their society "once and for all" (*TJ*, p. 176/153). I shall refer to this condition as the *perpetuity* condition.

In some places, Rawls distinguishes perpetuity from ultimacy and implies that the two conditions are imposed separately (*TJ*, pp. 147/146–47). When Rawls refers to "the condition of finality" in the *Kantian Congruence Argument*, I believe that what he has primarily in mind the *ultimacy* condition. But it is helpful to bear in mind the dual valence of "finality." Doing so makes it possible to locate another thread connecting the *Argument for Relative Stability*, the *Argument from Love and Justice*, and the *Kantian Congruence Argument*—a set of arguments which are generally read as self-contained but in which the first overlaps the second and the second is strengthened by the third. As we have seen, all three of these arguments concern the strains of commitment, for each concerns itself with the question of whether members of the WOS will be able to honor in principles of justice in perpetuity. Since the perpetuity constraint is imposed by finality broadly understood, the first two arguments therefore ask, in effect, whether members of the WOS will be able to honor principles chosen subject to the finality condition. Seeing this, we should recognize that the transition to the third argument, with its explicit appeal to finality, is less abrupt than it seems at first blush. Finality has been at stake all along.

11. For the importance of distinguishing the two conditions, see Freeman, "Reason and Agreement," p. 145, note 38.

12. The wording and context of a similar remark in "Reply to Alexander and Musgrave" suggest synonymy even more strongly; see Rawls, *Collected Papers*, p. 250.

There is another reason for bearing both senses of "finality" in mind when we read the *Kantian Congruence Argument*. We saw near the end of §VII.3 that someone who treats the principles as ultimate treats the desire to act from them as "regulative with respect to other desires" (*TJ*, p. 574/503). Someone who commits herself to treating the principles as final in that sense may therefore commit herself to treating the desire to act from the principles as supremely regulative over a considerable stretch of her future, since she will take steps to confirm herself as a just person and to keep unjust desires from arising or gaining too much force when they do. And so she may be said to treat the desire to act from the principles as regulative, not just of her desires, but of her plans. But there are firmer grounds for saying this if she also commits herself to treating the principles as final in the second sense, for she then treats the principles as holding in perpetuity, and hence for the duration of her life. If treating the principles as final in the first sense entails treating her desire to act from them as regulative of desires, treating them as final in the second sense entails treating the desire to act from them as regulative of the whole of one's future life.

I shall argue that finality is one of the conditions of the OP that makes it the case that our rational nature determines the choice there, and that this helps to explain Rawls's appeal to the finality condition in the *Kantian Congruence Argument*. If my argument is successful then—since we have seen why Rawls accepts the *KI Claim*—we will see why he accepts (5.2), understood expansively as claiming that the desire to express our nature as free, equal, and rational is the desire to act on principles chosen subject to the conditions of the OP, including the finality condition. Seeing why Rawls accepts (5.2), we will see how the *Kantian Congruence Argument* goes and why taking the desire to act from the principles as supremely regulative is good, as judged from within the thin theory. In making the argument, I shall generally take "finality" to refer to the *ultimacy condition* alone. There will, however, be some points at which Rawls's argument is illuminated by construing finality as including the *perpetuity condition* as well, since what is judged to be good is treating the principles as supremely regulative of one's plans into the indefinite future.

What, exactly, do ultimacy and perpetuity have to do with the good of expressing our nature as free and equal rational persons?

Rawls remarks at one point that "a person may be regarded as a human life lived according to a plan" (*TJ*, p. 408/358). On its face, this is a suggestive but odd remark. Whatever else Rawls means by it, the remark suggests that persons and plans are connected in such a way that we can learn some of what Rawls thinks about *persons* by learning what he thinks about *plans*. Rawls thinks we live as befits rational beings, and so express our nature as rational, when our plans of life are framed and pursued rationally. He also says that "the unity of the person is manifest in the coherence his plan" (*TJ*, p. 561/491). Thus if plans are rational and unified, persons are rational and unified. That is because when persons carry out rational and unified plans, they act as deliberators capable of ordering their ends, and they act as purposeful agents, rather than as beings torn by conflicts of desire they cannot resolve. Plans that

are framed and pursued rationally are plans that exhibit the unity of reason, or what Rawls calls simply "unity." Living lives that exhibit rational unity is one of the ends we have insofar as we are rational beings. Inasmuch as we are rational, we desire to live such lives.[13]

My suggestion is that our nature as beings who have this end determines our choice in the OP because of the finality condition. For, as we have seen those who acknowledge the principles as final treat them as supremely regulative. And Rawls thinks that plans of life can be unified in the right way only if they are framed and pursued in accord with principles of right that are treated this way. So the effect of imposing the finality condition is that our rational interest in unifying our practical reason helps to determine the choice in the OP. That is why, as (1.9) suggests, the OP is a choice situation in which our *rational* nature is one of the decisive determining elements. To make good on the suggestion, I have to say what Rawls means by "unity." That must be teased out of several scattered passages.

Unities of the Self

The remark that "the unity of the person is manifest in the coherence of his plan" occurs in *TJ*, §85, the section on "The Unity of the Self." Rawls continues immediately "this unity being founded on the higher-order desire to follow, in ways consistent with his sense of right and justice, the principles of rational choice" (*TJ*, p. 561/491). This suggests two features of a unified plan that are worth distinguishing for analytical purposes.

- *Dictation of Plans by Rational Choice*—A unified plan is one in which the ends a plan includes are chosen and scheduled according to the principles of rational choice. The contrast I believe Rawls has in mind is that between plans that are constructed and pursued on the basis of reasons, and those in which ends are adopted, balanced, and pursued on the basis of what he calls "purely preferential choice."[14]
- *Consistency of the Right and the Good*—The ends to which the person is committed at any one time should not, insofar as far as she can tell, be such that any one, once duly specified, precludes pursuit of another. The consistency with which Rawls seems especially concerned in the passage on unity is a consistency between her sense of justice and the various ends that are connected with her conception of the good. So someone has a unified plan only when the ends she includes in it are consistent with her judgments of what ends are just and unjust.

13. Here I rely on Rawls's tantalizing remark about regulative ends at *TJ*, p. 415/364.

14. Why consider this feature of a plan a feature that *unifies* it? Presumably, it is because among the reasons we have for including certain ends in our plans are that they fit with ends chosen before and their adoption and pursuit make sense in light of previous choices, so that the whole plan is unified over time.

Later, in *TJ* §85, Rawls notes that "judgments of rights are to be reasoned and not arbitrary" (*TJ*, p. 562/492). This suggests a third feature of unified plans.

- *Rationality of Right*—This feature requires that the principles of justice with which someone's ends must be consistent are such that they lead to judgments of right that the agent herself sees to be supported by good reasons.

Finally, earlier in *TJ*, speaking of the rational person's plan, Rawls says "the whole . . . has a certain unity, a dominant theme" (*TJ*, p. 420/369). This suggests still another feature of unified plans.

- *Unity of Character*: This is the kind of unity that would lead us to say of someone living out a plan which exhibits it that she has an identifiable character that shows itself in her deliberations[15] and that persists over time.

Rawls thinks we pursue rational plans—and so act as rational agents—when our plans exhibit these four features, and achieve them in the right way. And Rawls thinks our plans will achieve these four features in the right way if and only if they are regulated by principles of right. These are strong claims. I turn now to the ways Rawls would defend them. Those defenses show how we unify our agency by treating the desire to act from the principles as supremely regulative.

Regulative Principles and the Unities of the Self

Let's start with Rawls's reasons for thinking that framing our plans in accord with regulative principles is necessary if our plans are to have these four features in the right way.

The alternative to living out plans that accord with principles that we must take as supremely regulative is, Rawls thinks, for "each to draw up his rational plan without hindrance under full information" (*TJ*, p. 565/495). On the basis of these plans, each can lodge claims of other citizens and of institutions that are presumptively valid. Principles of justice then have the role of adjudicating conflicts among these presumptively valid claims. Let us call this the *Priority of Good Alternative.* The problem with the *Priority of Good Alternative* is that it implies a dilemma: either the plans of agents will not exhibit the four features of the unity of reason, or those plans will be unified but they will be unified in a way that is unacceptable because it deforms the self.

On the *Priority of the Good Alternative*, any ends are candidates for inclusion in a plan because there are no prior principles of justice to rule any out of bounds. This has unfortunate consequences. "How in general," Rawls

15. Recall that Rawls describes someone's "fundamental character" as "the ordering that determines the weight of reasons"; see Chapter VI, note 5.

asks "is it possible to choose among plans rationally?" (*TJ*, p. 551/483) "Using the principles of rational choice as guidelines," he continues "and formulating our desires in the most lucid form we can, we may narrow the scope of purely preferential choice, but we cannot eliminate it altogether" (*TJ*, p. 552/483). For "sooner or later," he says "we reach incomparable aims." To the extent that plans are determined by purely preferential choice among incomparable aims, they are rationally indeterminate: the principles of rational choice fail to single out one end rather than another for inclusion. At that point, Rawls says, "significant intuitionist elements enter into determining the good" (*TJ*, p. 560/491). To the extent that plans are rationally indeterminate, they lack the first element of rational unity, *Dictation of Plans by Rational Choice*. This is something of a difficulty because the failure of reason to guide our choice can leave us feeling "unsettled" (*TJ*, p. 450/395).

The indeterminacy of plans when all ends are open for consideration is not itself a *serious* difficulty. But the consequences of this indeterminacy can be, at least when any ends adopted by individuals are *ipso facto* the grounds of presumptively valid claims of justice. For the natural way—Rawls assumes the only way—rationally to determine what justice demands on the *Priority of the Good Alternative* is for "society [to] proceed[] to maximize the aggregate fulfillment of the plans that result" (*TJ*, p. 565/495). That is, the natural way rationally to determine what justice demands on the *Priority of the Good Alternative* is to embrace a teleological theory of justice. In that case, the indeterminacy of plans is problematic for "in a teleological theory any vagueness or ambiguity in the conception of the good is transferred to that of the right" (*TJ*, p. 559/490). Why is this problematic?

The plans whose aggregate fulfillment must be maximized according to the *Priority of the Good Alternative* are plans framed in part on the basis of purely preferential choice. So what principles of justice demand, according to this alternative, ultimately depends upon such choice. This flies in the face of our considered judgment that "what is right is not a matter of mere preference" (*TJ*, p. 559/490). It also means that plans that include satisfying the principles will not fully exhibit the third feature of rational unity, the *Rationality of the Right*. For suppose I must do without some resources because justice demands the satisfaction of others' ends, and suppose I know those ends to have been adopted on basis of others' purely preferential choice. Then the claims that are being honored in preference to mine will seem arbitrary. So too will the verdict of justice that supports honoring those claims. By raising the possibility that individuals may not be able to see the demands of the right as rational, it raises the possibility that individuals will not find reason to maintain their sense of justice. This possibility threatens the stability that congruence was supposed to help secure.

Still another problem posed by the rational indeterminacy of life plans can be seen if we consider the possibility that a majority favors repression of some religious practice. Rawls thinks there is "no sure way" to rule out such preferences as irrational (*TJ*, p. 450/395), so the end of repressing abhorrent

practices can be included in a rational plan of life. Principles of justice that require society to maximize the aggregate satisfaction of rational plans may therefore require the prohibition of these practices "even though they cause no social injury" (*TJ*, p. 450/395). Such a prohibition may strike us as arbitrary. In that case plans of life that include satisfying the principles will lack the *Rationality of the Right* for a second reason as well.

Furthermore, a shift of majority preferences or of preference intensities may mean that conduct which is permitted at some time may justly be prohibited or discouraged at another. Unified plans exhibit *Consistency of the Right and the Good*. This requires that agents' ends be consistent with the principles of justice. On teleological views, this requires that agents may have to sacrifice pursuit of their ends to whatever is demanded by the maximal aggregate satisfaction of citizens' plans. If what is demanded changes because of shifting preferences in the citizenry, ends that could once have been pursued may later have to be dropped from agents' plans so that *Consistency* is maintained. The content of plans is therefore always hostage to shifts in the preferences of others. The conditions needed for long-term planning are not secure. The liability of plans to change threatens the kind of long-run unity that I have said Rawls thinks is important, plans' *Unity of Character.*

In all these ways, the indeterminacy of plans of life on the *Priority of the Good Alternative* and the "transfer" of that indeterminacy to principles of justice threaten plans' rational unity. This conclusion reminds us of just how deeply Rawls is troubled by the possibility that reason will leave the demands of justice indeterminate—in which case the content of those demands could be affected by the "intuitionist elements" that inevitably enter into determining each person's good. I have commented elsewhere that readers generally ignore Rawls's concern with intuitionism because so much of his effort throughout *TJ* is devoted to defeating utilitarianism, and because they forget that utilitarianism itself seems attractive because it promises to avoid the problems with intuitionism.[16] In fact, the possibility that the demands of justice will be rationally indeterminate is a pervasive concern in *TJ*. Time and again, Rawls insists that while reliance on intuition and on purely preferential choice are not completely eliminable, it is a virtue of justice as fairness that it limits the affects they have on the principles of right (see *TJ*, pp. 41ff/37ff).

Justice as fairness avoids the difficulties that beset the *Priority of the Good Alternative* in what Rawls thinks is the only acceptable way: agents unify their lives by framing their plans in accord with principles which must be taken as ultimate and perpetual. And he thinks that framing our plans in accord with

16. Cf. "Classical utilitarianism tries, of course, to avoid the appeal to intuition altogether. It is a single-principle conception with one ultimate standard; the adjustment of weights is, in theory any way, settled by reference to the principle of utility.... Undeniably one of the great attractions of the classical doctrine is the way it faces the priority problem and tries to avoid relying on intuition" (*TJ*, p. 41/36).

such principles is part of the way we express our rational nature *because* he thinks framing our plans in accord with such principles is necessary to give our lives the unity of reason. Rawls's discussion of congruence and the unity of life therefore grows out of what I have identified as a pervasive concern of *TJ*: the concern that rational indeterminacy will affect the content of demands of justice.

To return to the argument against the *Priority of the Good Alternative*: I said earlier that Rawls thinks the *Alternative* faces a dilemma. We have already seen one horn of the dilemma—indeterminacy and the problems to which it leads. There is one way to eliminate the indeterminacy of life plans without demanding that agents frame their plans in accord with supremely regulative principles, but that way of eliminating indeterminacy is patently unacceptable. This is the other horn of the dilemma that the *Priority of the Good Alternative* faces. What is the unacceptable way of eliminating indeterminacy?

Rawls says at one point that the indeterminacy of plans when all ends are open for consideration "seems to arise, then, from the fact that a person has many aims for which there is no ready standard of comparison to decide among them when they conflict" (*TJ*, p. 552/484). So if there *were* some standard by which conflicting ends could be compared and ordered, then indeterminacy could be eliminated and the *Priority of the Good Alternative* would be viable. The only plausible such standard, Rawls argues, would be an end to which the conflicting ends are subordinate. With such an end in hand, agents can settle conflicts among ends by determining how best to pursue the superordinate end (cf. *TJ*, p. 552/484). Of course, if there is more than one superordinate end, then these could also conflict. That conflict would presumably have to be settled by appeal to an end that is superordinate to those ends. So on the *Priority of the Good Alternative*, indeterminacy can be eliminated only if agents adopt some one dominant end by which their conflicting ends can be compared and ordered. The problem with this way of eliminating indeterminacy is that it is "irrational, or more likely…mad" (*TJ*, p. 554/486) to treat all ends but one as the means to a dominant end. To behave this way would be radically to misvalue many of the good things in human life. So eliminating indeterminacy by treating one end as a dominant end is unacceptable. This is the second horn of the dilemma faced by the *Priority of the Good Alternative*.

How does justice as fairness avoid the dilemma?

Rawls insists that purely preferential choice is not completely eliminable from the formation of our plans of life. The first feature of unified plans, *Dictation of Plans by Rational Choice*, can be approximated (cf. *TJ*, p. 552/483) but it cannot be fully realized. But in justice as fairness, the demands of justice are given antecedently and regulate the choice of ends. The rational indeterminacy that is entailed by the ineliminability of purely preferential choice will not affect the right. This is evident from the ways in which other elements of unity are attained on a view that gives priority to the right.

Framing plans in accord with supremely regulative principles is obviously sufficient for the second feature of unified plans, *Consistency of the Right and*

the Good. For the person whose plan accords with supremely regulative principles takes the principles of justice as regulative of the ends his plan includes and the claims he makes on the basis of those ends. In that case, "desires and aspirations are restricted from the outset by principles of justice which specify the boundaries that men's systems of ends must respect" (*TJ*, pp. 31/27–28). Unjust preferences are taken to have "no merit in the first place" (*TJ*, p. 31/28) and ends cannot be included in someone's plan if they conflict with his sense of justice. Thus, insisting that agents act from supremely regulative principles is a way of coping with the fact—noted above—that the adoption of unjust preferences cannot be shown to violate the principles of deliberative rationality.

The third feature of unified plans is the *Rationality of the Right*. Plans that accord with supremely regulative principles do not face the difficulties with this feature I identified earlier. For supremely regulative principles do not make the right "a matter of mere preference" (*TJ*, p. 559/490) and they do not give any weight to unjust preferences.

Clearly a life-plan does not exhibit the *Unity of Character* just in virtue of its being permanently in accord with supremely regulative principles. But being framed in this way contributes to and facilitates this feature of unified plans. Someone who acts from supremely regulative principles of justice over a complete life is a consistently just person. In Chapter VI, we saw how a sense of justice transforms the person who has it, so that she consistently attaches greater weight to certain goods than the unjust person does. This consistency of valuation gives her life some unity. Moreover, when principles of justice are taken as supremely regulative of everyone's ends in perpetuity, the threats to the *Unity of Character* that arise on the *Priority of the Good Alternative* are avoided. Agents can live out their long-term plans with security.

Thus when plans are in accord with supremely regulative principles, the problematic implications of indeterminacy for the right are avoided. Some features of a unified plan are realized and others are facilitated. We might express this conclusion by saying that while rational unity of plans cannot be completely attained, "the *essential* unity of the self" (*TJ*, p. 563/493, emphasis added)—the unity that *makes* a self *a* self—is sufficiently provided for when agents live their lives in accord with supremely regulative principles.

Of course, expressing the conclusion this way depends upon the connection Rawls asserts between persons and plans. It is because "a person may be regarded as a human life lived according to a plan" (*TJ*, p. 408/358) that we can identify the conditions of unified rational *agency* by identifying the conditions of unified rational *plans*. Having identified the conditions of rationally unified plans, the connection among persons, plans, and supremely regulative principles seems more plausible. For if human beings do indeed have a rational nature which is realized in our actions, then that nature is surely what is common to us all insofar as we live rationally. Our nature is shown by what is common to rational plans of life or—as Rawls puts it—"the nature of the self as a free and equal moral person is the same for all, and the similarity in the

basic form of rational plans expresses this fact" (*TJ*, p. 565/495). The principles of justice are what give rational plans their "similarity in . . . basic form." It is the desire to act from supremely regulative principles of right that is common to all rational plans. That is why the principles "reveal our nature" (*TJ*, p. 560/491). It is because the principles "reveal our nature," that Rawls claims at a crucial point in the *Kantian Congruence Argument* that the desire to act from those principles "reveals what the person is" (*TJ*, p. 575/503).

Authorship of Our Plans

Living a unified life is an end we have inasmuch as we are rational and, inasmuch as we are rational, we desire that end. But what I have said so far about the unity of life is compatible with the thought that plans are simply, as it were, handed to people who live them out and whose primary interest in the unity of those plans is the interest they have as rational executors. This is surely not an adequate description of what the desire for a unified life is a desire for. Citizens of a WOS do not just want to be the *executors* of plans. They realize their nature, not just by executing rational plans, but by framing them. If they have a desire to express their nature, then that desire must be a desire to be the *framers* of their own plans or, as I shall say, their *authors*. They think of themselves as free and, thinking of themselves as free, they want to be authors of a particular kind. They want to be authors of long-term plans that they draw up by following reasons as they see them, rather than plans that are determined by "the contingencies and accidents of the world" (*TJ*, p. 575/503).

The claim that they have this desire is confirmed by the fact that an interest in free, long-run authorship explains the satisfaction they take in realizing various kinds of unity. We have already seen how treating principles as final, in the sense of "perpetual," helps to secure *Unity of Character*. An interest in that sort of unity reflects an interest in living according to plans which extend over the long run. Interest in the *Consistency of the Right and the Good* and in the *Rationality of the Right* do not themselves indicate an interest in being the author of one's plan of life. But the fact that citizens are satisfied with the way justice as fairness achieves these forms of unity does reflect such an interest. For in justice as fairness, *Consistency of the Right and the Good* is attained by treating principles of right as supremely regulative. *Rationality of the Right* is attained by singling out principles of right that are chosen by citizens' rational representatives in the OP. Since those are principles that citizens can regard as "self-imposed" (*TJ*, p. 13/12), citizens can see themselves as the authors of the reasonable principles that regulate their plans of life.

Citizens' interest in being the authors of their own plans is clearest from Rawls's treatment of the remaining kind of unity: *Dictation of Plans by Rational Choice*. Rawls thinks that the uncertainty or indeterminacy that infects our plans because of the ineliminability of purely preferential choice may leave us feeling "unsettled" (*TJ*, p. 450/395). The experience of feeling unsettled is mitigated in a society in which principles of justice satisfy the perpetuity condition.

For in such a society, in which everyone has long framed plans in accord with the same principles, generations of just citizens before us have tried out and validated various forms of life. "Thus in drawing up our plan of life, we do not start de novo; we are not required to choose from countless possibilities without given structure or fixed contours…the priority of justice securely constrains these deliberations so that they become more manageable" (*TJ*, pp. 563–64/494).

When deliberations are more manageable, there is no occasion for the completely unsettled feeling of having to choose a plan of life completely on one's own. If Rawls's view does not enable us fully to achieve the *Dictation of Plans by Rational Choice*, it leaves less scope for purely preferential choice than does the *Priority of the Good Alternative*. Thus, a WOS eases the job of freely authoring a plan of life and does so without imposing plans on citizens. Furthermore, because everyone observes "fixed contours," agents can decide for themselves how to live within them without fear of undue interference by others. I believe Rawls thought that what he calls this "free play" (*TJ*, p. 566/496) of practical reason is itself experienced as satisfying, at least when it is appropriately constrained by principles of right. And so I believe the Rawls of *TJ* thought that citizens of a WOS would find security and satisfaction in the distinctive ways justice as fairness copes with the ineliminability of purely preferential choice.

Thus the Rawls of *TJ* thought citizens of a WOS have a rational interest in being the free authors and executors of their own unified plans of life. This interest is part of their nature as rational. As we have now seen, Rawls argues that that interest is satisfied when and only when we conform to an ideal according to which life plans are regulated by principles of right. And so members of the WOS, including Joan, have a rational interest in living according to supremely regulative principles. Persons who acknowledge the finality of principles—understood now as both ultimate and perpetual—treat them as supremely regulative, so including finality among the conditions of the OP brings it about that our interest in living according to supremely regulative principles helps to determine choice there. That is one of the reasons Rawls moves from the *KI Claim*, via (1.9), to a critical step in the *Kantian Congruence Argument*, (5.2).

Seeing why Rawls makes this move, we can complete the sketch of the *Kantian Congruence Argument*. For we can now see why Rawls accepts

(5.2.1) We can satisfy the desire to act from principles chosen in the OP only if we treat that desire as supremely regulative.

And since the sense of justice is a desire to act on principles chosen in the OP, we can see why he accepts:

(5.5') Joan can satisfy the desire asserted in C_4a by and only by treating her sense of justice as supremely regulative of her other desires.

With (5.5') in hand, Rawls can get to:

(5.9') "The desire to [treat our sense of justice as supremely regulative of our other desires] and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire."

And we have seen that from (5.9'), Rawls can move—via the ostensible conclusion—to:

(5.12) Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans.

We have also seen that how short a step it is from (5.12) to *TJ's Nash Claim*:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

We have already seen how the *mutual assurance problem* can be solved for the WOS, so that Rawls can move from C_N to:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans.

Finally, establishing C_6 shows that desires to be unjust are outweighed by other desires members of the WOS have, quite apart from their desire to be just. Even if they feel tugs toward injustice, those tugs are countervailed. Furthermore, the viewpoint of full deliberative rationality differs from that of the thin theory only in that someone adopting the former point of view takes her desire to be just as such into account. If members of the WOS know that it is rational to maintain their sense of justice even when they leave that desire out of account, they know that it is rational to maintain it when they take account of it along with the rest of their desires. And so Rawls can move from C_6 to the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans.

Members of the WOS can draw up their plans knowing that they and others have more to gain by being just than not.

There is no one place in *TJ* where Rawls lays out the *Kantian Congruence Argument* from start to finish. Seeing the connections among the finality condition, the unity of the self and our nature as rational beings helps us to fill in details so that we can see just how the argument runs. The *Kantian Congruence*

Argument depends on a prior argument—reviewed in §IV.2—from the *Two Conjunct Reading* of the Aristotelian Principle to C_4a. The *Kantian Congruence Argument* itself then runs from C_4a—via the *KI Claim*, the treatment of the unity of the self and (1.9)—to (5.2) with its implicit imposition of the ultimacy and perpetuity conditions. (5.2), (5.2.1), and (5.2.2) get Rawls to (5.5'). The importance of publicity and transparency are asserted to justify (5.6) and the move from (5.5') to (5.7'). From (5.7'), the argument runs—via (5.9'), the ostensible conclusion and the relevant *Balance Conditional*—to (5.12). And from (5.12), it moves to *TJ's Nash Claim*, C_6, and the *Congruence Conclusion.* When fully laid out, the *Kantian Congruence Argument* is one of the most intricate but also, I believe, one of the most elegant, arguments in all of *TJ*.

Since the desire to live as a free and equal person would be encouraged and reinforced by the institutions of the WOS, it gives members of the WOS a reason to live justly that is not only decisive, but enduring. We saw in Chapter VI that they can reflect on their sense of justice at any point in life. The *Kantian Congruence Argument* is supposed to show that any time they do so, they would maintain it and they would not be pulled hard enough by temptation to accede. Of course, Rawls notes that "in practice all social systems are subject to disturbances of some kind" (*TJ*, p. 457/400). Should members of the WOS act unjustly, "principle guilt" leads them to make amends and restore equilibrium (*TJ*, p. 474/415). Rawls anticipates this point in "The Sense of Justice," where he says that "should . . . violations [of principles of right] nevertheless occur, in cases of temptation, feelings of guilt will tend to restore joint activity."[17] The WOS would therefore be, not only just, but stably just.

§VII.7: Kantian Unity

The conception of unity at work in the argument that culminates in C_C is a Kantian conception of unity—or at least, it is a conception of unity that Rawls ascribes to Kant. In his lecture on Kant entitled "The Unity of Reason," Rawls says that according to Kant, reason is unified when the legitimate claims of the two points of view within reason are honored and when neither power of reason overreaches or exceeds its authority. Rawls is talking there about how theoretical and practical reason are unified, but his lecture suggests how unity might be achieved within practical reason.

Recall that in §II.3, I said that congruence is a relation that obtains between two points of view—the viewpoint of full deliberative rationality and what Rawls calls "the standpoint of justice" (*TJ*, p. 569/498)—and that practical reason is unified when congruence obtains. Rawls's lecture on Kant suggests that the two points of view within practical reason are unified when the legitimate claims of each are secured and neither oversteps its bounds. Unity of this sort is just what is achieved when C_C is true and the principles of justice are judged to be

17. Rawls, "Sense of Justice," *Collected Papers*, p. 106.

final from within full deliberative rationality, for to treat the principles as final is to treat them as the final arbiters of competing claims. There is to be no checking of the result because all moral considerations relevant to the conflict have already been taken into account. The good has, as it were, already had its proper say in the determination of the principles. Rawls implies that the indeterminacy of the right on teleological views shows that "the structure of teleological doctrines is radically misconceived: from the start they relate the right and the good in the wrong way" (*TJ*, pp. 560/490–91). We might add: the indeterminacy shows that by relating them in the wrong way, they allow the capacity for a conception of the good to overstep its proper bounds.

Achieving the four kinds of unity insures that agency is unified both diachronically and synchronically. Since the ideal of rational unity is one that members of the WOS, such as Joan, must live up to if they are to live up to their *free-and-equal self-conception*, this suggests that diachronic and synchronic unity answer to something important in that self-conception. I believe that this suggestion is right.

Persons with a *free-and-equal self-conception* thinks of themselves, perhaps implicitly, as single agents who persist through time, with projects and plans which continue, but which they may also change while remaining the same persons. It is this view of themselves that accounts for their rational interest in achieving rational unity of agency or what Rawls calls "unity of the self." It is also this view of themselves that helps to account for the interest in freedom that parties in the OP honor on their behalf. This is especially clear in Rawls's treatment of liberty in the revised edition of *TJ* (see *TJ*, rev. ed., pp. 131–32). It is also at work in one of Rawls's early treatments of religious liberty where Rawls writes:

> one should ask what rational individuals in the original position could acknowledge as principles to regulate the liberties of the citizen. In this case, it is equally clear that they can acknowledge only an equal liberty of conscience and that this initial position must be final. If each person is thought to regard himself as in general subject to religious obligations (although he may expect that these obligations will change over his life if his religious views change), then he can only acknowledge the principle of equal religious freedom[.][18]

Spelling out the *free-and-equal self-conception* in the way that I now have may seem to suggest that Rawls is doing just what I said in §II.1 that he is not doing: beginning with a view of the person that presupposes conclusions about personal identity that have to be established by the disciplines of metaphysics or philosophy of mind. But the Rawls of *TJ* would insist that this conclusion is a mistake. He would insist that the *free-and-equal self-conception* is a *self-conception.* It is simply a way that members of democratic societies, including the WOS, think of themselves, perhaps implicitly.

18. Rawls, "Constitutional Liberty and the Concept of Justice," *Collected Papers*, p. 89. This essay dates from 1963.

Rawls does not provide philosophical arguments to show that this self-conception is the best way for members of such societies to think of themselves. Even in *TJ*, he simply takes the *free-and-equal self-conception* as a starting point, and asks what conception of justice would best settle conflicts among persons who think of themselves this way. His answer is, of course, justice as fairness—in part, as we have now seen, because taking the desire to act from the two principles as supremely regulative answers to the interests and desires the *free-and-equal self-conception* carries with it. In Chapter VIII, we shall see that Rawls came to recognize that this starting point was not as uncontroversial as he thought at the time he wrote *TJ*.

§VII.8: Korsgaard, Unity and the Bridge Function

As I have read the *Kantian Congruence Argument*, Rawls would defend (5.2) by drawing on a line of thought that runs parallel to a line of argument Christine Korsgaard has explored in her Locke Lectures and elsewhere. Yet there seems to be an important difference between my argument and Korsgaard's, a difference that may cause doubt about whether I have interpreted the *Kantian Congruence Argument* correctly.

To see the parallel, recall that on my reading, Rawls assumes that people think of themselves as unified agents who persist through time, and that they have a rational interest in acting as such. Rawls argues that we are unified by maintaining the desire to act from principles of right as supremely regulative. So we have a rational interest in maintaining those principles as regulative and satisfying that interest is part of our good. The question is what those principles are. That is the question choice in the OP is supposed to answer. The finality condition of the OP insures that our interest in maintaining principles as supremely regulative helps to determine choice of principles. And so acting from principles of right chosen in the OP is part of our good.

Korsgaard's Locke Lectures[19] and related papers are too rich and interesting to be examined here. I mention just one of her arguments because of what it shows about Rawls's use of the OP. Very roughly, Korsgaard argues that if we are to act, then we must see ourselves, or think of ourselves, as unified agents.[20] Taking the categorical imperative as supremely regulative of our action unifies us. That, she argues, is why categorical imperatives have normative force for us—that is "the way they bind us."[21]

Suppose we take seriously Rawls's suggestion that principles of justice "are" or "are analogous to categorical imperatives" (*TJ*, p. 253/222)—and

19. As of this writing, Korsgaard's Locke Lectures are available through links on her web site: http://www.people.fas.harvard.edu/~korsgaar/

20. See Christine Korsgaard, "Personal Identity and the Unity of Agency: A Kantian Response to Parfit," *Philosophy and Public Affairs* 18 (1989): pp. 101–32.

21. Christine Korsgaard, *Locke Lecture II*, "Self-Constitution: Action, Identity and Integrity," p. 1.

hence that categorical imperatives, or principles analogous to categorical imperatives, would be adopted in the OP. Then we can see why the line of thought I have read into the *Kantian Congruence Argument* runs parallel to the argument from Korsgaard's Locke Lectures, and we might be unsettled by the fact that Korsgaard follows her line to a very different conclusion. Perhaps, we will think, the fact that principles chosen in the OP unify the self is supposed to explain why those principles are normative for us, *not*—as I have maintained—why acting from them is satisfying or is part of our good.

But Korsgaard's argument should not, I think, raise doubts about my interpretation of the *Kantian Congruence Argument*. Rawls designed the OP to play what I have called the *bridge function*. It is supposed to bridge the right and the good. More precisely, since the various conditions that define it are chosen because they impose commonly accepted condition on arguments about principles of right (*TJ*, p. 18/16) *and* because they bring it about that our nature is the "decisive determining element" of agreement, the same conditions that make the chosen principles obligatory also make acting from them part of our good. Principles chosen subject to one of those conditions, namely finality, are such that acting from them unifies us. Since that condition is part of what suits the OP to play the *bridge function*, it should not be surprising if this consequence can be exploited both to show—as Korsgaard, in effect, argues—that principles which would be chosen in the OP *bind* and to show—as Rawls argues—that acting from those principles is *good*.

§VII.9: Is the OP Necessary?

I now want to return to an important and much controverted question I raised in §I.3: the question of whether the OP is essential to Rawls's development of his theory of justice. I said then that this question arises because some readers have thought it possible to offer a sound argument for Rawls's two principles without appealing to the OP. The availability of such an argument would not, of course, entail that the OP is not essential to the theory, since it is possible to that the OP plays an essential role elsewhere. The fact that Rawls designed the OP to play the *bridge function* suggests places where it might play an essential role. Even if the OP is not essential to the argument for the two principles, it might be essential to an argument or a set of arguments for congruence. Before I ask whether it is, I need to look into what is meant by saying that the OP is not essential. I can do so most clearly by referring to the Pivotal Argument.

Those who think it possible to defend Rawls's principles without appeal to the OP do not, I think, deny that principles of justice must be acceptable to those to whom they apply. And so they would agree to one of the critical steps of the Pivotal Argument, the step which says that:

(1.6) The principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal persons.

The Pivotal Argument builds on this step to move to:

(1.10) The principles governing the ways the basic structure distributes primary goods must be acceptable in the OP.

and to:

C_1: The distribution of primary goods by the basic structure must be governed by the two principles.

because of the way it interprets what is required for principles to be acceptable to us "as free and equal persons."

Those who deny that the OP is essential accept (1.6), but they think it is possible to move from (1.6) to C_1 without appealing to the OP because they deny that the demand imposed by (1.6) requires that principles be acceptable to us in the OP. Rather, they think acceptability under some other set of conditions is enough. That set could be a subset of the conditions that define the OP, or it could include some but not all of the conditions that define the OP together with some additional conditions. It could, for example, include the formal constraints on the concept of right, but admit fuller information than the veil of ignorance allows. Of course, if the set does include the formal constraints, then those constraints would need to be motivated independently of their connection with the OP. But if those constraints really are, as Rawls says, "natural" and "weak" (*TJ*, p. 131/113), then motivating them should not be a problem. What matters for present purposes is that those who deny that the OP is essential think the requirement imposed by (1.6) can be satisfied by acceptability under a *different* set of conditions than those that define the OP, so that Rawls could have defended the principles without appeal to their acceptability in *it*. And so those who deny that the OP is essential would move from (1.6) to C_1 without going through the intervening steps in the Pivotal Argument.[22]

As I indicated in §I.3, the best attempt to show that the OP is nonessential of which I am aware is that by Joshua Cohen. Cohen thinks Rawls's two principles can be defended by showing that they would be accepted at every social

22. In §I.3, we saw that according to Ronald Dworkin, Rawls argues for his two principles "through" the OP. What he means is that the OP "enforce[s] the abstract right to equal concern and respect," which he takes to be "the fundamental concept of Rawls's deep theory." I said then that Dworkin is sometimes thought to have shown that the OP is not essential to the argument for the principles. It should now be apparent that this conclusion is a mistake. Dworkin thinks, I believe, that "the abstract right to equal concern and respect" requires principles of justice to be acceptable to us as persons. I therefore think that Dworkin accepts (1.6). Whether or not the OP is essential to justice as fairness depends upon whether it is necessary to appeal to all the conditions that define the OP in order to move from (1.6) to C_1. In Dworkin's terms, it depends upon whether acceptability under those conditions is *necessary* "enforce the abstract right to equal concern and respect." Dworkin's arguments do not settle that question.

position.[23] Cohen's argument depends upon some of the formal constraints, but not on the veil of ignorance. He concludes, in effect, that the OP is not essential to the argument for C_1.[24]

Even if it is possible to argue for C_1 without appealing to acceptability in the OP, the fact that the OP plays the *Bridge Function* raises the question of whether Rawls could defend the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

in the spirit of *TJ* without appealing to the OP.

23. If Cohen's argument does what I have claimed can be done—namely, move from (1.6) to the principles—it must be because showing that principles are acceptable at every social position is enough to show that they are "acceptable to us as persons." But what reason is there to think that?

Let me sharpen the question. Rawls thinks that he can show his two principles are acceptable to us as persons by showing that they would be accepted in the OP. In *TJ*, he thinks that because he works with a conception of persons according to which (1.9) is true. Thus, he can appeal to that conception of the person to argue that principles must be accepted by parties veiled in ignorance and subject to the other constraints of the OP. But now suppose it is granted, at least for purposes of argument, that the OP is objectionable because the veil allows too little information. So suppose it is granted that if the acceptability requirement imposed by (1.6) is to be satisfied, the principles must be chosen in light of more information than the veil of ignorance admits. I have read Cohen as, in effect, claiming that the requirement can be satisfied by requiring that principles be accepted by choosers who know what they would know if they occupied each social position. But, it might be objected, what principled reason is there for stopping *there* with *that* information? We are trying to enforce the requirement that principles be acceptable to us *as persons*. That seems to demand that the amount of information allowed contracting parties be justified—as in Rawls's arguments—by a conception of the person. But what conception of the person could justify stopping anywhere between the veil of ignorance and full information?

The answer is that the amount of information on the basis of which principles are adopted in Cohen's argument is not dictated by a conception of the person. It is dictated by the nature of the choice contracting parties must make. I have spoken of the "acceptability" of principles for simplicity's sake. In fact what needs to be shown is not just that Rawls's principles are "acceptable," but that they would be chosen from a menu of principles after a series of pair-wise comparisons. The principles on the menu differ only in how members of the WOS at different social positions would fare if they were adopted, so contracting parties do not need more information than Cohen allows them. Thus for my purposes, the beauty of Cohen's argument is that it shows how it is possible to enforce (1.6) while remaining neutral among conceptions of the person. This neutrality will prove extremely important; see Chapter VIII, note 33.

24. The conclusion that the OP is not essential is explicit at Cohen, "Democratic Equality," p. 751; Cohen explicitly appeals to one of the formal constraints—finality—at p. 736.

It may seem hard to know how to address this question, because it seems hard to say what Rawls *could* say or what arguments would or would not be in the spirit of *TJ*. So to sharpen the question, I want to ask:

> Could Rawls have defended C_C using a strategy that is recognizably like that he used to defend it in *TJ*, but without appealing to the OP?

As we have seen, Rawls's strategy has three steps.

> First, he establishes that there are certain relevant desires that all members of the WOS normally have—those referred to by C_4a, C_4b, C_4c, and C_4d.
>
> Second, he argues that those desires can best or only be satisfied by realizing ideals which require treating the sense of justice as supremely regulative.
>
> Third, he shows that, given the special circumstances of the WOS, this provides decisive reasons for treating the desire to act from principles of justice as supremely regulative.

As we saw in Chapter V, the arguments at the second step depend upon what I called the *diversity of descriptions*. They depend, that is, upon the fact that contract theory "supplies [a diversity of] descriptions of what the sense of justice is a desire for" (*TJ*, p. 569/499). The various descriptions depend upon conditions that define the OP, such as finality and publicity. So it may seem doubtful that the answer to my question is "yes."

But showing that the OP is not essential to the defense of some conclusion does not require producing an argument for that conclusion that does not appeal to *any* of the conditions that define the OP. It requires producing a sound argument that does not appeal to *all* of them. In addition to producing an argument for C_1 that appeals to some but not all the conditions that define the OP, Cohen tries to establish congruence by appeal to some but not all of those conditions. So Cohen may seem to promise an affirmative answer to my question after all.

Cohen notes, however, that his congruence argument is limited in two ways. First, his argument is not an argument for C_C. It is an argument for the weaker conclusion that the congruence of the right and the good is more likely in the WOS of justice as fairness than in a society well-ordered by competing principles. Second, Cohen limits himself to an argument from the desire referred to by C_4d—to an argument from what I called "the desire to participate in forms of life that call forth talents."[25] Rawls says, correctly I think, that while those who have this desire have reason to maintain the desire to act from the two principles as supremely regulative, the argument that this reason is decisive appears sound only if we make a simplifying assumption that is

25. The limits are stated at Cohen, "Democratic Equality," p. 749; the competing principles of justice with which Cohen is concerned are not utilitarian but those that are part of "mixed conceptions" of justice.

somewhat dubious. That is why he thinks the *Argument from Love and Justice*, which moves from C_4b, C_4c and C_4d to C_C, needs to be supplemented. So the argument Cohen produces does not purport to show, nor could it to show, that the conclusion of Rawls's congruence arguments C_C can be defended using Rawls's strategy but without appeal to the OP.

What argument could show this? The argument for C_C with which Rawls supplements the *Argument from Love and Justice* is the *Kantian Congruence Argument*. That argument starts with C_4a, which asserts that members of the WOS have a desire to express our nature. As we saw in §VII.4, the argument gets from C_4a to C_C via what I called the *KI Claim*:

> 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*, p. 253/222)

and:

> (1.9) The OP is a choice situation in which our nature is the decisive determining element.

At the beginning of this section, I proposed what I hope is a helpful way of thinking about the thesis that the OP is not essential to the defense of Rawls's principles of justice. Those who deny that it is essential grant the first half of the Pivotal Argument, and so they grant the requirement that principles must be acceptable to us as persons. But they think that the conclusion of that argument C_1 can be reached without going through (1.9), because the acceptability requirement can be satisfied by showing that the principles would be accepted under a different set of conditions than those that define the OP.

This suggests a way of thinking about whether the OP is essential to Rawls's treatment of congruence. Grant C_4a, and with it the claim that members of the WOS want to express their nature. And grant the *KI Claim*, which equates expressing one's nature with acting from principles that would be accepted under conditions in which one's nature determines the choice. Then ask whether it is possible to get from C_4a and the *KI Claim* to the conclusion C_C while bypassing (1.9), because there is a different set of conditions than those that define the OP in which our nature determines what principles are accepted. If some such set of conditions can be identified, then it is possible to defend the *Congruence Conclusion* using the strategy Rawls employed in *TJ*, but without appealing to the OP. It would follow that, as the OP is shown not to be essential to the argument for Rawls's two principles by Cohen's argument, so it is not essential to the argument for congruence either. There would then be a strong case for thinking the OP is not essential to *TJ*.

It is not possible to go through all imaginable conditions and ask whether they would do. It is, however, very instructive to see why the OP cannot be shown to be non-essential by a move that the success of Cohen's arguments tempts us to make. It may be tempting to suppose that:

(1.9*) If principles are acceptable to persons regardless of what social position they occupy, then the nature of persons must be what determines that the principles are acceptable.

It seems to follow from (1.9*) and the *KI Claim* that members of the WOS express their nature by acting from principles that are acceptable to persons in every social position. If so, then acting from principles that are acceptable from every social position satisfies the desire referred to by C_4a, the desire to express our nature. And since Cohen's argument shows that Rawls's two principles are acceptable from every social position, it follows that members of the WOS express their nature when they act from Rawls's principles. So if the rest of the *Kantian Congruence Argument* is valid, then—if (1.9*) is right—it seems possible to move from C_4a to the conclusion of Rawls's congruence argument C_C without relying on (1.9) or appealing to the OP.

But (1.9*) is *not* right, at least if social positions are distinguished only by income, wealth, and opportunity (cf. *TJ*, p. 96/82). For even if principles are acceptable to persons regardless of their social position, it could still be that those principles are acceptable to persons—singly or collectively—*because of their conception or conceptions of the good*. And since there is no one conception of the good or set of conceptions that it is natural for people to endorse, conceptions of the good are not part of our nature.[26] And so what would determine the acceptability of principles in that case is not our nature after all.

Someone who wants to show that the OP is not essential to the congruence argument could avoid this difficulty with (1.9*) by relying on:

(1.9**) If principles are acceptable to persons regardless of what social position they occupy *and what conception of the good they endorse*, then the nature of persons must be what determines that the principles are acceptable.

But while Cohen's argument shows that Rawls's two principles are acceptable to every social position—and so satisfy the antecedent of (1.9*)—his argument does not show or purport to show that Rawls's principles satisfy the antecedent of (1.9**). Moreover, because of the endless variety of conceptions of the good, it is very hard to see what argument *could* show this except an argument which showed that the principles would be accepted by persons who did not know which conception out of that infinite variety they actually endorsed. It is therefore very hard to see what argument could show that Rawls's principles satisfy the antecedent of (1.9**), except for an argument that shows that those principles would be adopted by persons subject to the informational constraints imposed by the veil of ignorance.

Thus an argument that tries to move from C_4a and the *KI Claim* to the *Congruence Conclusion* C_C, while bypassing (1.9) by relying on (1.9**) instead,

26. See Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, pp. 549–50.

will still have to show that Rawls's principles are acceptable to persons who are subject to the OP's most salient defining condition. Perhaps they will not have to show that the principles are acceptable to persons who are subject to *all* of the conditions that define the OP. Perhaps some will be omitted or revised. But analysis of the *Kantian Congruence Argument* showed that the argument from C_4a to C_C appeal to finality and publicity. It is therefore not clear what conditions of the OP might be dispensed with.

Moreover, the informational poverty of the OP is the condition that critics find most objectionable. The alleged objectionability of the informational constraints of the OP is what motivates some of Rawls's defenders, including Cohen, to show that the OP is not essential to Rawls's arguments for the principles. But if those constraints are essential to an argument *from* C_4a and the desire to express our nature *to* C_C, as I have contended, then it is hard to see what would motivate an attempt to show that the *OP* is not essential to such an argument. And if an argument from C_4a is needed to establish C_C because of shortcomings in the *Argument from Love and Justice*, as Rawls rightly says, then it is hard to see what would motivate the attempt to show that the OP is not essential to the Rawls's treatment of congruence.

To help determine whether the OP is essential to justice as fairness, I asked:

> Could Rawls have defended C_C using a strategy that is recognizably like that he used to defend it in *TJ*, but without appealing to the OP?

That is the question with which I have been concerned in this section. The considerations I have now brought forward strongly suggest that the answer to this question is "no." The Rawls of *TJ* may be able to defend his two principles by an argument that resembles the Pivotal Argument, but that bypasses (1.9) and the OP. But he cannot establish C_C and congruence without them. To that extent, at least, the OP is essential to the theory.

§VII.10: Conclusion

With the completion of the *Kantian Congruence Argument*, *TJ*'s discussion of congruence is fully before us. The argument concludes *TJ*'s treatment of inherent stability. I indicated in Chapter II just how ambitious that treatment is. If it succeeds, Rawls has shown how terms of cooperation that are collectively rational can be individually rational as well, as judged from within both full deliberative rationality and the thin theory of the good. Terms of cooperation we would give ourselves can, when institutionalized, elicit a desire to act from them. When institutionalized, they also form us so that we see that acting from those principles expresses our nature as we understand it. We become people who value acting from principles we give ourselves. In this way, the terms of cooperation, when institutionalized, can remove the threat of

collective action problems and stabilize themselves in a large, modern society. Showing that they can would itself be a tremendous accomplishment.

The arguments for inherent stability depend upon the educative or formative work of just institutions in a WOS. The arguments succeed only if those institutions bring about members' enduring convergence on certain views of themselves and their freedom, and on certain ends, such as the end of living up to ideal-dependent desires and the end of living as free and equal rational beings. The arguments will succeed only if those institutions foster effective motives to pursue those ends. Of course, institutions could not do their formative work if human beings were naturally incapable of acquiring these ends or of effectively pursuing them. If the arguments for inherent stability seem plausible, then we must find it plausible that our nature is amenable to such moral formation.

The claim that it is runs counter to a powerful and recurrent strain of argument in the West, a strain developed and transmitted by those Rawls refers to as "the dark minds in Western thought." Prominent among these are Augustine and Dostoevsky. While Hobbes's view of humanity is not as dark, his work bears relevant affinities to theirs.[27]

This is hardly the place to sketch the complex ethical and social views of these three thinkers.[28] What matters for present purposes is that the three of them seem to have held that human beings are too sinful, weak, and self-interested to live under free and just political institutions. Indeed, Augustine seems to have thought that political institutions were inherently coercive and punitive, and that they would have been unnecessary if human beings had remained in the state of innocence in which God created us.

Augustine's, Dostoevsky's, and Hobbes's views of human nature find clear expression in what they said about how political societies need to be stabilized. Augustine thought that society could be just only if it was unified by worship of the true God. Since all earthly societies are composed of members of the City of God and the City of the World, no such society is just. Every earthly society is a modus vivendi in Rawls's sense, stabilized—to the extent that it is—by mutual knowledge of each person's desire for peace. The Dostoevsky of the *Grand Inquisitor* seems to have thought that human beings would be frightened and confused by freedom, and that society must be stabilized by religious authoritarians. Hobbes, of course, thought terms of cooperation would be undermined by collective action problems unless the terms were stabilized by an absolute sovereign.

27. See Chapter II, note 26. I am grateful to Ronald Beiner for pointing out important differences between Hobbes on the one hand, and Augustine and Dostoevsky on the other.

28. For an elementary introduction to Augustine, see my "Augustine's Political Thought," *The Cambridge Companion to Augustine* (Cambridge: Cambridge University Press, 2001), ed. Eleonore Stump and Norman Kretzmann, pp. 234–52. The description of Augustine's views in this and the following paragraphs follows that essay, and sources are cited there.

These views of human nature are potent threats to liberal democracy, since those who accept them will think liberal democracy an unworkable ideal—if they think it an ideal at all. By identifying a form of political freedom that we can sustain, and by showing that a WOS can be inherently stable, Rawls hoped to vindicate a different and brighter view of our nature. The idea of reciprocity lies at the heart of that view. The *bridge function* played by the OP—the fact that the OP connects the right and the good—makes its centrality clear. Fair terms of cooperation are, Rawls would say later, "terms that each participant may reasonably accept, provided that everyone else likewise accepts them" (*PL*, p. 16). The conditions defining the OP guarantee that the terms adopted there are fair, and are therefore terms of what we might call "reasonable reciprocity." Those conditions also guarantee that human beings living in a WOS express our nature as free and equal rational beings when we regulate our lives by those terms. And so, on Rawls's view, living together on terms each thinks others could accept, being treated fairly and responding in kind, all suit and express our nature as it would unfold under just institutions. The *Kantian Congruence Argument* appeals to this claim to show that a WOS would be stably just. The claim expresses Rawls's view of what kind of lives our nature suits us to lead. By showing what we can be, Rawls hoped to ground reasonable faith in human beings and in the real possibility of a just, liberal and democratic society.[29]

But Rawls's argument for the inherent stability of justice as fairness depends, as we have seen, on a solution to the *mutual assurance problem*. While I said in §VII.3 that the problem would be solved in the WOS by common knowledge and readily observable features of justice institutions, this assurance is not complete, and it cannot be. Crimes and cheating can go unreported or underreported. Even if differences seem to be adjudicated by a shared conception of justice, for all each knows, there may still be a significant number of people who do not attach sufficient weight to living as free and equal moral persons to judge that they should treat their sense of justice as supremely regulative. If members of the WOS are to regulate their plans by Rawls's two principles, they will need some assurance that the defectors will either obey the principles under duress, or will not prove so disruptive that principles of nonideal theory come into play. And so, as I noted in §II.2, Rawls concedes that the WOS will have to rely on sanctions to ensure compliance and to clinch the *mutual assurance problem*. He says, "Of course, under normal conditions public knowledge and confidence are always imperfect. So even in a just society it is reasonable to admit certain constraining arrangements to insure compliance[.]" (*TJ*, p. 577/505)

This concession may seem to pose a serious problem for Rawls. The reliance on sanctions may seem considerably to reduce the distance between Hobbes's approach to stability and Rawls's, and to diminish the distance

29. I spell out this line of thought at greater length in "John Rawls and the Task of Political Philosophy."

between imposed stability and the kind of stability Rawls said would be enjoyed by justice as fairness. To determine whether this is right, it is necessary to see exactly what the sanctions in the WOS are supposed to do. Rawls says of them "their main purpose is to underwrite citizens' trust in one another. These mechanisms will seldom be invoked and will comprise but a minor part of the social scheme" (*TJ*, p. 577/505). I take him to mean that C_6 is true, and that everyone in the WOS judges that it is good to be just, but—because members have limited assurance that C_6 is true—sanctions would have to be attached to defection to make it clear to everyone that no rational person would choose to defect.

The facts that the congruence arguments are successful and that C_6 is true show that justice as fairness has done much to stabilize itself, by bringing it about that members of the WOS want civic friendship and the goods of social union, and want to live as free and equal persons. This in itself marks a significant difference from Hobbes's view, since Hobbes seems to have thought that people would judge it in their interest to comply with terms of cooperation only when the cost of sanctions are taken into account.

A further difference with Hobbes emerges when we recall that Rawls wanted to show the inherent stability of justice as fairness because of the connection between inherent stability and autonomy. The principles of justice are principles members of the WOS would give themselves. To show inherent stability is to show that institutions that conform with those principles bring it about that those who live under them give them their informed and willing support. It is to show, we might say, that members of the WOS would act from, and not just in accord with, the principles they would give themselves. This is what is really significant about inherent stability.

In Hobbes, at least as Rawls reads him, stability is imposed because there is a clear sense in which sanctions are imposed. As we saw in §II.1, Rawls thinks "the Hobbesian sovereign is . . . an agency *added to* an unstable system of cooperation in such a way that it is no longer to anyone's advantage not to do his part given that others will do theirs."[30] The phrase "added to" is important. Members of a Hobbesian society establish a sovereign who alters their payoff tables without facing any such payoff table himself, since he is not a player in the prisoner's dilemma game.[31] His decisions about what sanctions to impose are immune from appeal and moral constraint. By contrast, in the WOS, the system of sanctions would not be imposed in these ways. It would be administered by officers of the WOS who are themselves subject to the sanctions, who are subject to principles of justice, and who must respect the rule of law (cf. *TJ*, p. 576/504). Thus in the WOS, unlike Hobbes's society, there are clear ways in which even the system of sanctions is self-imposed, and thus clear ways in which the stability they help to bring about is inherent. It is possible to

30. Rawls, "Sense of Justice," *Collected Papers*, p. 104 (emphasis added).

31. See Ullmann-Margalit, *Emergence of Norms*, p. 64.

avert the "hazards of the generalized prisoner's dilemma" without installing a Hobbesian sovereign (*TJ*, p. 577/505).

Unfortunately, despite the ingenuity of the *Argument from Love and Justice* and the *Kantian Congruence Argument*, Rawls came to believe that *TJ*'s case for inherent stability was fatally flawed. With the arguments for congruence before us, I can pinpoint the premises and arguments he rejected, and show why he made the transition from *TJ* to *PL*. I shall begin to do that in Chapter VIII.

VIII

The Great Unraveling

At the beginning of Chapter II, we saw that Rawls said he made the changes between *TJ* and *PL* because *TJ*'s account of stability "is not consistent with [his] view as a whole." That account is an account of *inherent* stability. Rawls wanted to show that justice as fairness, when implemented and publicized, would stabilize itself. *TJ*'s account of stability falls into two parts. In the first, Rawls argues that the institutions of the well-ordered society (WOS) would encourage members' effective desire to be just. In the second, he argues for the congruence of the right and the good. I have said that Rawls took his political turn because he became dissatisfied with *TJ*'s treatment of congruence. Now that we have seen exactly how the arguments for congruence go, it is time to see why Rawls came to find them unsatisfactory. It will be useful to recall where we have come so far and to anticipate the interpretation to be defended in this chapter.

Rawls's principles of justice can, of course, be represented as the object of a collective agreement on terms of social cooperation. But if members of the WOS think they have something to gain by "free-riding" on the justice of others, they will be tempted to defect from that agreement. In that case, the stability of justice as fairness will be threatened by "the hazards of the generalized prisoner's dilemma" (*TJ*, p. 577/505). To show that the hazards would not undermine agreement on the terms of cooperation, the Rawls of *TJ* tried to show that each person in the WOS would judge, when making her plans from her own point of view, that she is better off being a just person than she would be if she were the kind of person who decided whether to act justly or

to free-ride case-by-case. More precisely, he tried to establish what I called the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

If the stability that results is to be inherent rather than imposed, the argument for C_C cannot depend upon the presence of a Hobbesian sovereign or a dominant religion. In Chapter III, we saw that one route to C_C starts from the claim that just institutions encourage all members of the WOS to live up to certain ideals. It starts, that is, from:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship and association included in justice as fairness.

or at least from:

C_3*: All members of a WOS want to live up to the ideal of full autonomy.

Rawls says that the question of whether the right and the good are congruent arises in trivial and nontrivial forms, depending upon what desires members of the WOS are assumed to have. Prisoner's dilemmas are trivially averted if the prisoners all have and know they all have an effective desire to cooperate. The hazards of the generalized prisoner's dilemma can be trivially averted in the WOS if everyone has and knows that everyone else has effective desires for objects the value of which depends upon the value of being just—that is, desires for objects the value of which is given by what Rawls calls the "full theory of the good." C_3 says that everyone in the WOS has such desires. An argument for C_C that begins with C_3 or C_3* answers only the trivial form of the congruence question.

Prisoner's dilemmas really arise when we first suppose that prisoners lack an effective desire to cooperate, and then ask what they have reason to do from what Rawls calls the "self-interested point of view." The "hazards of the generalized prisoner's dilemma" threaten the stability of justice as fairness when we put aside everyone's desire to be just, and the related desires referred to by C_3 and C_3*, and imagine that the typical member of the WOS, whom I called Joan, adopts—not a self-interested point of view—but the point of view associated with what Rawls calls the "thin theory of the good." The thin theory is an account of value that does not presuppose the value of being just for its own sake. When Joan adopts this point of view, she values being just—and living up to the ideals of justice as fairness—only to the extent that they get her other things she wants.

I have argued that Rawls thought there are some desires that all adult members of the WOS would normally see that they have when they follow the thin theory of the good. Those desires are encouraged by just institutions, and are referred to by the conclusions established in Chapter IV:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

In Chapters VI and VII, we have seen how the Rawls of *TJ* tries to move from C_4a, C_4b, C_4c, and C_4d to the conclusion that each member of the WOS would judge—even from within the thin theory of the good—that being just is her best reply to the justice of others. This is the conclusion I called *TJ's Nash Claim* and expressed as:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

C_N expresses a judgment members of the WOS would reach whenever they reflect on their sense of justice and ask whether they maintain it. If the arguments for C_N succeed, then—since the *mutual assurance problem* can be solved—the threat of the generalized prisoner's dilemma is averted and Rawls can infer:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

Since Rawls can get to C_C from C_6, members of the WOS regulate their plans by their sense of justice. The reasons they do so are enduring, and justice as fairness is "as stable as one can hope for" (*TJ*, p. 399/350). Because the sense of justice, the truth of C_N, and the solution to the *mutual assurance problem* all result from the institutions that implement and publicize justice as fairness, the stability justice as fairness enjoys is inherent stability.

The ideals referred to by C_3 and C_3* are ethical ideals. To suppose members of the WOS would converge on those ideals is to suppose that they would converge on what Rawls later called a "partially comprehensive doctrine" (*PL*, p. xviii). In the years after he published *TJ*, Rawls came to recognize that such convergence is unrealistic, even with the encouragement of just institutions, because those institutions would also encourage ethical pluralism. It is because Rawls later read *TJ*'s account of stability as depending upon this convergence that he described *TJ*'s account of stability as "unrealistic" in *PL* (*PL*, p. xviii).

The realization that C_3 and C_3* are unrealistic raised a question of stability that was not taken up in *TJ*:

> If some members of the WOS do not have the ideal-dependent desires implied by C_3 or C_3*, is it rational for them to maintain their sense of justice on the basis of the various comprehensive views of the good they *do* hold?

If those conceptions of the good converge on the desires referred to by C_4a, C_4b, C_4c, and C_4d, and if the only or the best way to satisfy those desires is by being just when others are just, then Rawls could defend an affirmative answer to *this* question, using the *Argument from Love and Justice* and the *Kantian Congruence Argument*. But while Rawls did not deny that they converge on those desires, he came to think that the arguments from C_4a, C_4b, C_4c, and C_4d fail. Indeed, he came to think C_3 and C_3* are unrealistic *because* he came to think those arguments fail, for the arguments Rawls offered for C_3 and C_3* depended upon claims in the congruence arguments that he came to regard as untenable. The new question of stability had to be answered in a new way: by appeal to an overlapping consensus.

In this chapter, I shall begin to substantiate this reading of the changes between *TJ* and *PL*. First, I want to look more closely at the content of the ideals to which C_3 and C_3* refer. One of those ideals, of course, is full autonomy. I remarked a moment ago that in the *Dewey Lectures*, Rawls argues that members of the WOS will normally develop a desire to realize this ideal. But while the ideal is not explicitly appealed to in §86 of *TJ*, I noted in §III.2 that Rawls does refer to this ideal in other sections of *TJ*. The presence of the ideal there facilitated Rawls's retrospective reading of C_3 or C_3* into *TJ*. A closer look at the content of the other ideals justice as fairness includes shows why Rawls could later maintain that those ideals, too, are presupposed in *TJ*'s account of stability.

§VIII.1: The Content of Ideals

I said Rawls argues that members of the WOS can best or only satisfy the desires referred to by C_4a, C_4b, C_4c, and C_4d by conforming or living up to certain distinctive conceptions of conduct, friendship, and association. That means that the arguments from C_4a, C_4b, C_4c, and C_4d make it possible to fill in the content of the ideals referred to by C_3.

The argument from C_4c, for example, turns on the fact that even someone who follows the thin theory takes the desire to act from the principles as supremely regulative because only that way can she be sure of protecting her friends and the social forms she cares about. But in daily life, members of the WOS do not follow the thin theory; they follow the full theory. They act from

the principles because their ties of affection extend widely, and they see that the content of the principles makes the principles suitable norms for the mutual justice that friendship requires. Friendship, including civic friendship, among those who are moved by considerations of justice is the ideal of friendship that justice as fairness includes.

The argument from C_4d turns on the fact that someone who takes the desire to act from the principles as supremely regulative thereby participates in a social union of social unions. This suggests that in daily life, members of the WOS treat the principles as supremely authoritative because they see that—in virtue of the content of the principles—basic institutions that conform to them encourage diversity and guarantee each the resources she needs to make use of her liberties. The conception of a social union is the ideal of association that justice as fairness includes. The ideal of a social union of social unions—while not strictly speaking an ideal of association, since the social union of social unions is not a *voluntary* association—shows how important features of this ideal are realized in the WOS itself.

These ideals of friendship and of a social union of social unions have clear implications for the ideal of personal conduct. More details of that ideal can be filled in by looking at the other congruence arguments.

The argument from C_4b turns on the fact that someone who takes the desire to act from principles of justice as supremely regulative conducts herself according to principles she can openly avow before others. To avoid what I called a "justification gap," she uses those principles to explain herself when necessary, so that others have assurance that she acknowledges those principles. Of course, insofar as members of the WOS follow the thin theory, they have reason to conduct themselves in these ways because they want to avoid the costs of hypocrisy and deception. But in daily life, they are moved by the content of the principles. They act from the principles because they see that their content suits them for regulating the conduct of persons who want to act from principles they can openly avow. Being the sort of person who takes the content of the principles as a reason to act from them is part of the ideal of personal conduct in justice as fairness.

The argument from C_4a, as elaborated by the *Kantian Congruence Argument*, supplies even more detail. Someone who preserves her sense of justice as regulative lives a life that is unified at a time and over time in various important ways. She avoids the division of self that comes with hypocrisy and deception. She has a character, her good is consistent with the demands of right, and her plan of life is rational. According to the *Kantian Congruence Argument*, someone following the thin theory who takes the desire to act from the principles of right as supremely regulative enjoys what I called "thin autonomy". But in daily life, she enjoys full autonomy. This is not only because she acts from principles that she knows would be adopted in the OP—and hence would be adopted freely—but also because of the content of the principles from which she acts. Principles with that content are such that, when institutions conform to them, no one's plan of life is dictated by contingencies

that have an illegitimate bearing on their choices. When the principles are satisfied, everyone's plan of life is the work of free practical reason. This fact about the content of the principles is part of what moves someone who realizes full autonomy to act from them.

The ideal of personal conduct unifies the *diversity of descriptions* by showing how pursuit of the various ends a sense of justice is a desire for can be combined into a single life. This fact will prove important in §IX.2 when we ask why Rawls developed his definition of a sense of justice between *TJ* and *PL*. I argued in §III.2 that this ideal, and the other ideals of justice as fairness, are not metaphysical conceptions. They are, however, ethical ones. By arguing that their realization belongs to everyone's good in the WOS, Rawls was arguing that members of the WOS would converge on a partially comprehensive doctrine. But what makes the ideals to which C_3 refers *ethical* ideals?

Someone who conducts her friendships justly, participates in a social union of social unions, and realizes full autonomy treats principles of right as finally authoritative. But she cannot realize those ideals if she treats the principles as regulative only of the claims she makes in political life. She realizes them by treating those principles as regulative of *all* of her conduct and practical deliberation. Moreover, the realization of those ideals depends upon her recognizing that those principles are ones she would give herself freely, in the OP. Rawls states this explicitly in the case of full autonomy, saying that it is realized in "maintaining the first principles that would be adopted in [the OP] and publicly recognizing the way in which they would be agreed to."[1] And so the person who realizes the ideals justice as fairness includes does so in part because she knowingly treats principles she would give herself—rather than divine commands or natural law, for example—as ultimate moral principles. That is what C_3 says belongs to the good of members of the WOS.

But if Rawls really did suppose that all members of the WOS shared a comprehensive conception of the good, and that conception was justice as fairness itself, why did he think that justice as fairness was a *liberal* view? Hasn't Rawls himself taught us that in order to be liberal, a conception of justice must accommodate the fact of pluralism?

It would be natural to reply by pointing to the priority of liberty, to the importance of resources in giving liberties their equal value, and to the place of autonomy within justice as fairness. But if what is being asked is why Rawls would have thought a view that assumes convergence on ethical ideals is appropriate for a pluralistic society, then I think the answer starts elsewhere.

TJ treats justice as fairness as *comprehensive*, so that its ideals apply to the whole of life. But it is *partial*. It does not, for example, include detailed ideals for family life, nor does it say which private association or social union members of the WOS are to join. One can live up the ideals while choosing a wide variety of occupations, vocations, and leisurely pursuits. Rawls says emphatically in *TJ* that

1. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 315.

justice as fairness "does not aim at a complete specification of conduct" (*TJ*, p. 566/496). So convergence on the ideals of justice as fairness is, Rawls thought, compatible with choosing a wide variety of ways of life.

This conclusion is supported by the fact that the ideals of justice as fairness are *higher order*. They do not concern the content of fully comprehensive doctrines. Rather, they concern the ways in which people conduct the friendships they have, and the ways in which they come to and hold whatever fully comprehensive doctrines they embrace. Rawls thought, I believe, that someone could hold a wide variety of religious or secular views of life while conforming to the ideal of personal conduct. What conformity to the ideal demands is that one's fully comprehensive view of the good, whatever it is, be arrived at and held as the work of free practical reason. Just what this requires is suggested by the discussion of the unity of the self in Chapter VII. Conceptions of the good should be arrived at on the basis of reasons. If their precepts of right are not to be rationally indeterminate, then anyone holding those conceptions must still treat the principles of justice as finally authoritative. Some of the conceptions embraced in the WOS may be religious. Those who hold them can still realize the ideal of full autonomy provided they hold those views in the right way.[2]

Now that we have given a bit more content to the ethical ideals to which C_3 refers, it seems clear that the ideals of friendship and association—the ideals of civic friendship conducted according to the principles, and of a social union of social unions—are in *TJ* along with the ideals of a unified self and of full autonomy. The ideal of a social union of social unions is clearly present and, indeed, a whole section is devoted to it. The ideal of just friendship also seems clearly to be present once we know where to look.

What is less clear, as we saw in §III.2, is that the Rawls of *TJ* thought members of the WOS had ideal-dependent desires or that those desires contribute to stability in *TJ*. In §III.2 I said that there is some textual basis for these claims in *TJ*. But even if *TJ* does not include arguments for C_3 or for the presence of ideal-dependent desires, it does suggest how members of the WOS might develop such desires. They could develop them by reflecting on a fact that the congruence arguments reveal: the fact that conducting friendships justly, participating in a social union, and realizing full autonomy satisfy certain of their natural desires—the desires referred to by C_4a, C_4b, C_4c, and C_4d. This in itself should show that living up to the ideals is attractive and desirable. The attraction of the ideals is heightened when they see that the principles of

2. I believe Rawls thinks there is no other way to avoid the rational indeterminacy of religious ethics; see *TJ*, pp. 554/485–86. Rawls's argument presupposes what seems to be correct: that reliance on religious authority to give revelation determinate content threatens a justification gap. But Rawls may also think that the precepts of religious ethics would need the validation of reason even if they were not indeterminate. On this, see his instructive contrast between Kant and Leibniz at *Lectures on the History of Moral Philosophy*, pp. 229–30.

justice and their applications—from which they all benefit—are explained by reference to the ideals.

If this suggestion is right, then it is a fairly short step from the congruence arguments of *TJ* to the presence of those desires. Once the step is taken, it is clear how ideal-dependent desires contribute to congruence and stability. We can then see why Rawls would later have implied that *TJ*'s treatment of stability assumes C_3, and why he would have thought that the failure of the congruence arguments in *TJ* undercuts support for C_3. My suggestion about how ideal-dependent desires are developed gets some confirmation from the way Rawls later said members of the WOS would develop the desire to be fully autonomous.

§VIII.2: Defending C_3

The ideal-dependent desire for full autonomy is a desire to be a certain sort of person. In §III.2, I said that Rawls came to think that desire was part of the sense of justice, yet *TJ*'s extensive discussion of how members of the WOS acquire a sense of justice gives little indication of how this desire develops. The presence of the desire depends upon someone's having and wanting to live up to a conception of what she can be. How do members of the WOS acquire this view of themselves? And how do they develop the desired to live up to it?

To see the answer, let's return to the *Kantian Congruence Argument*. The first premise of that argument is C_4a. The argument therefore assumes that Joan has a *free-and-equal self-conception*, as (1.1) says. She thinks of herself as a free person and she wants to express her nature as such. Step (5.7') of the *Kantian Congruence Argument* implies that Joan's desire to express her nature moves her to treat her sense of justice as supremely regulative. It is on the basis of (5.7') that Rawls infers (5.9'), the claim that the desire to treat our sense of justice as supremely regulative and the desire to express our nature specify what is practically speaking the same desire. But as we saw, these two desires will move Joan in the same way only if she has a conception of herself as naturally free and equal, and—insofar as she follows the thin theory—would want to express her nature by realizing a particular kind of freedom: thin autonomy.

Thin autonomy is an ingredient of the full autonomy members of the WOS realize in everyday life when they act from the principles; it is contributed by the fact that the principles were freely chosen in the OP. The *Kantian Congruence Argument* assumes that Joan grasps and would value this conception of her natural freedom because of what she learns about herself from the public conception of justice. To see this, recall that in the WOS, the publicity condition is satisfied—as Rawls asserts between steps (5.5) and (5.6)—and that Joan has a "lucid grasp" of justice as fairness. Having this grasp, she knows that her basic institutions treat her as the kind of being described in (1.1):

(1.1) We are by nature free and equal rational agents who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

She grasps what I called the *KI Claim*: "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*, p. 253/222). She also knows:

(1.9) The OP is a choice situation in which our nature is the decisive determining element.

And so she knows:

(5.2) The desire to express our nature is a desire to act on principles that would be chosen in the OP.

It is because Joan knows (5.2) that she knows she expresses her nature as free—knows that her conduct realizes a kind of freedom that is natural to her—by acting from principles adopted in the OP. Since expressing her nature is part of her good, so too is acting from the principles. Insofar as Joan follows the thin theory, she cannot value following the principles because of their content. But knowing all that she does when the publicity condition is satisfied, she *does* value the freedom she realizes in virtue of the freedom with which the principles would be adopted. In sum, as a result of what she knows by publicity, she values *thin autonomy*.

Thus in *TJ*, the role of the OP in the adoption of the principles—because publicly known—is assumed to have an educative effect. It affects how members of the WOS think of their nature as free and equal rational beings. Because they want to express their nature as free, it affects the kind or conception of freedom that they value. Rawls thought more deeply about the educative affect of the publicity condition in the years following the publication of *TJ*. This had profound implications for his account of moral education and more specifically, I believe, for his account of how members of the WOS develop and maintain their sense of justice. To see that, we need to look at what Rawls says in the *Dewey Lectures* about how members of the WOS acquire the desire for full, rather than thin, autonomy.

At the end of the second of the original *Deweys*, Rawls says:

> Once [the full publicity condition] is imposed, a moral conception assumes a wide role as part of political culture. Not only are its first principles embodied in political and social institutions and public traditions of their interpretation, but the derivation of citizens' rights, liberties and opportunities invokes a certain conception of their person. In this way, citizens are made aware of and educated in this conception. They are presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain. Thus the realization of the full publicity condition provides the social milieu within which the notion of full autonomy can be understood and within

which its ideal of the person can elicit an effective desire to be that kind of person.[3]

When Rawls says "once [the full publicity condition] is imposed . . . the derivation of citizens' rights, liberties and opportunities invokes a certain conception of their person," I take him to mean that the public justification of the principles and their application appeals to a conception of their person as free and equal, reasonable and rational. When Rawls says that "in this way, citizens are made aware of and educated in this conception," I take him to mean that public availability of the argument for, and application of, the principles encourages citizens to think of themselves as being free, equal, reasonable, and rational in the ways that that public conception of the person says they are.

Rawls is at pains to stress that the OP represents members of the WOS in these ways. Since the OP is part of the justification of the principles, I take it that appeal to the OP plays an important role in educating citizens in the relevant conception of themselves. It can play this role because members of the WOS know the *KI Claim*, (1.9) and (5.2), and see the connection between acting on principles of right and expressing their nature—just as Rawls maintained in the *Kantian Congruence Argument*. In the *Deweys* as in the *Kantian Congruence Argument*, members of the WOS see the connection between their nature as free and the freedom with which the principles are chosen. They therefore see (5.7') and (5.9').

But that is not all they see. They also see a connection between their nature and the *content* of the principles. They see that the content of the principles is appropriate to regulate a society of persons who are naturally free because when basic institutions satisfy the principles, members of the WOS choose their conceptions of the good freely. Knowing that their nature is represented in the OP and knowing how the principles were chosen there, they can then see that they were adopted in the OP to advance their natural interest in freedom. That is why Rawls says that "the original position . . . serves to connect, in the most explicit possible manner, the way the members of the well-ordered society view themselves as citizens with the content of their public conception of justice."[4] When members of the WOS regulate their lives by principles, the content and derivation of which they know to be appropriate for beings who are free, equal, reasonable, and rational, they realize full autonomy.[5] Knowing that they realize full autonomy, in turn, affects the way they think of their own natural freedom. They think this is the kind, or one of the kinds, of freedom that is natural to them to realize. That is why Rawls says that "the realization

3. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 339–40.

4. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 339.

5. Recall that full autonomy is realized by members of the WOS in their daily lives by "maintaining the first principles that would be adopted in [the OP] and by publicly recognizing the way in which they would be agreed to, as well as by acting from these principles as their sense of justice dictates." See Rawls, "Kantian Constructivism in Moral Theory," p. 315.

of the full publicity condition provides the social milieu within which the notion of full autonomy can be understood."

We have seen that Rawls thinks members of the WOS have a natural desire to express their nature as free. When I defended the *Two Conjunct Reading* of the Aristotelian Principle, I argued that what this desire is a desire for is responsive to what members of the WOS believe about their nature. So when they want to express their nature as free, what they want depends upon what kind of freedom they think is natural to them. Thinking that full autonomy is natural to them, then, members of the WOS want to live as fully autonomous persons. This desire is further encouraged when members of the WOS see that all benefit by the work of just institutions, and when those benefits are justified by public appeal to the principles and their derivation from the ideal of the person. Thus the presence of the ideal of the fully autonomous person in the public culture of the WOS "elicit[s] an effective desire to be that kind of person." This is an ideal-dependent desire.

The argument that members of the WOS would develop an ideal-dependent desire for full autonomy thus depends upon the assumption that members of the WOS think of themselves as naturally free and equal, as (1.1) asserts, and so have a *free-and-equal self-conception*. It depends upon their having a desire to express their nature as free and equal, as C_4a asserts. It also depends on the assumption that the *KI Claim*, (1.9), (5.2) are all publicly known and affect the way members of the WOS think of the kind of beings they are. Finally, as I have reconstructed it, the argument depends upon the assumption that members of the WOS know and can reflect on (5.8) and (5.9'), and so can see how they satisfy their desire to express their nature by taking the desire to act from principles as supremely regulative.

When Rawls says that members of the WOS have a "lucid grasp" of justice as fairness, he cannot mean that they are able to reconstruct the *Kantian Congruence Argument* in its entirety. But if they are to acquire a desire for full autonomy, they must have at least an implicit grasp of its main points. Even this may seem unrealistic, but it is not. We citizens of contemporary liberal democracies think of ourselves as bearers of rights who live in a way that is natural and congenial to us when the exercise of those rights is protected and enjoyed. The conception of ourselves that we have is one into which we have been educated by our social and political institutions, by our political culture and public documents, and by our public practices of justification. To say that we have this conception of ourselves is not to say that we have a nuanced grasp of, say, a theory of rights; it is merely to say that we have an implicit grasp of the main conclusions that such a theory would support. Rawls believes that members of a WOS would have a similar grasp of justice as fairness.

By the time of the *Deweys*, Rawls had come to think that ideal-dependent desires are part of the sense of justice. He also recognized that they are important for congruence, and that eliciting them is one of the ways that institutions generate their own support. These latter points are clear from methodological

remarks about justification earlier in the *Deweys*.[6] There Rawls assumes, I think, that a conception of justice will be stable only if it is mutually acknowledged, and that it can be mutually acknowledged by members of the WOS only if they can justify the conception to one another. He then says that the conditions for justifying a conception of justice depend upon there being a basis for citizens' political reasoning and understanding within public culture.

If no such basis is readily available—as Rawls thinks there isn't—then philosophy can identify and propose one. It may discover and formulate principles from the public culture that can serve as the basis of agreement. Or it may "originate and fashion starting points for common understanding by expressing in a new form the convictions found in the historical tradition." This, of course, is what justice as fairness does. It "propose[s] . . . conceptions and principles congenial to [the] most essential convictions and traditions" of a democratic society. For example, the conception of the person and the ideal of full autonomy that justice as fairness proposes are congenial to liberal democracy because they specify the *free-and-equal self-conception* that is held by members of liberal democratic societies. Thus to find a basis for agreement, philosophy may have to propose conceptions and ideals that are novel. Citizens might be "presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain."[7]

But the hope is that once proposed, this conception or ideal of themselves will fit with their deepest considered judgments. When this ideal is accepted and publicized, it enters public culture. There, it provides a public basis for discussion and justification. As we have seen, if the educational work of publicity is successful, members of the WOS develop a desire to live up to the ideal of full autonomy. If they then see that they can satisfy that desire only by preserving their sense of justice as supremely regulative, they can see that the good and the right are congruent, and stability is enhanced. As if to confirm the connection between congruence and the ideal of the person proposed by justice as fairness, Rawls says that "what justifies a conception of justice is . . . its *congruence* with our deeper understanding of ourselves and our aspirations, and our realization that, given our history and the traditions embedded in our public life, it is the most reasonable doctrine for us."[8]

6. In this paragraph and the following two, I draw on—and occasionally paraphrase—Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, pp. 305–7.

7. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 340.

8. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, pp. 306–7 (emphasis added). See also *PL*, p.312, where Rawls imagines someone who denies that liberty of conscience is a basic liberty. Rawls says "One way to continue the discussion is to try to show that the scheme of basic liberties as a family is part of a coherent and workable conception of justice appropriate for the basic structure of a democratic regime and, moreover, a conception that is *congruent* with its most essential convictions" (emphasis added).

The argument that the institutions of the WOS would elicit a desire for full autonomy is not, of course, enough to establish

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

For the argument does not show that members of the WOS would develop ideal-dependent desires to participate in a social union of social unions or to maintain ties of civic friendships based on justice. But the argument that they would develop a desire for full autonomy suggests how Rawls could argue that they would develop these other desires, and hence how he could defend C_3.

Rawls could argue that members of the WOS know that they desire to live as friends with others, as C_4c says. When the publicity condition is satisfied and members of the WOS have a "lucid grasp" of justice as fairness, they would see that the content of the principles of justice suits them to regulate relations of civic friendship. They also know that they have a desire to take part in social forms that elicit talents, as C_4d says. They would see that talents are elicited and developed in the WOS because the basic institutions of the WOS satisfy the principles of justice. They would also see that, because they help to uphold the principles, they themselves have a role in bringing forth those talents. They would know that they find it satisfying to live in a society in which the capacities of human nature are so widely developed. They would also know:

(4.5') "When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept" (*TJ*, p. 523/459).

They could therefore connect this satisfaction with the justice of their society, and the security it provides them to develop and exercise their own talents. They could also connect the justice of the WOS with the availability of friendships that they find satisfying. Seeing these connections, Rawls could argue, elicits ideal-dependent desires to conduct just friendships and to participate in a social union of social unions. If living up to these ideals *both* satisfies everyone's sense of justice *and* belongs to the good rational persons would choose in the WOS, then the right and the good are congruent.

In Chapter III, I noted that Rawls thinks *TJ* treats justice as fairness as a comprehensive moral view, and that what makes a moral conception comprehensive is its inclusion of various ideals of personal conduct, friendship, and association. I then argued that what he came to think was unrealistic about his early treatment of stability was that it assumed members of the WOS all want to live up to those ideals. And I suggested at the end of §VIII.1 that these ideals are all in *TJ*. That is why I read Rawls as thinking that his early treatment of stability depended upon C_3. But if Rawls's early treatment of stability did depend upon C_3, then Rawls must have had some arguments for the presence of the other ideal-dependent desires—arguments like those I just sketched.

When I ask why Rawls would have found C_3 unrealistic, it will be important that one of those arguments—like the argument that members of the WOS would have a desire for full autonomy—depend upon a premise from one of the congruence arguments in *TJ*. In this case, the premise is (4.5').

The arguments for these other ideal-dependent desires are not, however, to be found in Rawls's texts. Despite what is suggested by his own remarks about what makes a doctrine comprehensive, perhaps it will be said that Rawls did not think he ever relied on C_3. But even if he did not, the *Deweys* show that Rawls did think all members of the WOS would normally acquire an ideal-dependent desire to be fully autonomous, and that he therefore accepted:

C_3*: All members of a WOS want to live up to the ideal of full autonomy.

In what follows, I shall assume for the most part that Rawls thought his early treatment of stability depended upon the stronger C_3. But whether he did—or instead thought it depended upon the weaker C_3*—the explanation of the changes between *TJ* and *PL* remains basically the same. For C_3 entails C_3*, and both assert that members of the WOS would converge on ethical ideals.

In Chapter III, I also conjectured that Rawls came to think the ideal-dependent desires referred to by C_3 are central to the sense of justice. There are passages in *TJ* that anticipate the centrality of these desires. A clearer indication that Rawls came to hold this view can be found in an essay published just four years after *TJ*, "The Independence of Moral Theory." There, Rawls says that "a basic form of moral motivation is the desire to be and to be recognized by others *as being a certain kind of person*."[9] The context of this remark indicates that the moral motivation he has in mind is the desire to be just. If my conjecture is right, then the arguments I have imputed to Rawls in this section imply a significant addition to *TJ*'s discussion of moral development, since *TJ* gives little indication of how ideal-dependent desires are to be acquired. But because I do not think that the evolution of Rawls's thought about the sense of justice and its development were complete until *PL*, I shall postpone discussion of it until Chapter IX.

I have said that Rawls made the changes he did between *TJ* and *PL* because he concluded that C_3—or C_3*—was unrealistic. And he concluded that C_3—or C_3*—was unrealistic because he came more deeply to appreciate the pluralism of the WOS. But it would be a mistake to think that pluralism about the good straightforwardly implies this conclusion, or that the inconsistency Rawls found in justice as fairness was simply an inconsistency between C_3—or C_3*—and the fact of pluralism. As I pointed out in §VIII.1, Rawls thought that C_3 was consistent with members of the WOS holding a wide variety of conceptions of the good. We need to see why Rawls ceased to believe that. In this section, I have tried to show that Rawls's arguments for C_3—or C_3*—depend upon members of the WOS grasping the central claims of the congruence arguments offered in *TJ*. To see why Rawls came to think that C_3—or C_3*—

9. Rawls, "Independence of Moral Theory," *Collected Papers*, p. 293 (emphasis added).

was unrealistic, we need to see why a deeper appreciation of pluralism led him to think that those arguments for congruence failed. That will be the task of the next two sections. Having seen why Rawls thought pluralism undercuts those arguments, we will then be able to see how it undercuts the arguments for C_3—or C_3*. Only then will we see what inconsistency Rawls found within the original presentation of justice as fairness.

§VIII.3: Pluralism and the Failure of Congruence

When Rawls said that *TJ* "fails to allow for the condition of pluralism to which its own principles lead,"[10] he gave pithy expression to the upshot of a long and important argument. To see how pluralism undercuts the congruence arguments of *TJ*, we need to understand just how Rawls thought it arises in a free society. Recall that according to the *Public Basis View*, Rawls simply assumes the fact of pluralism. In §I.6, when I was looking at the philosophical difficulties with the *View*, I insisted that this is a mistake and that Rawls thought deeply about how free institutions encourage pluralism. The explanation of pluralism is especially clear in the case of the WOS, and I shall concentrate on that case. I think the account can be extended to cover free institutions that are not ideally just, but I shall not attempt that extension here.

Let's grant Rawls that members of the WOS realize, and know they realize, full autonomy in daily life. They realize it because their society is just, and the justice of its institutions insulates them from "social [and natural] contingencies" (*TJ*, p. 73/64). Their plans of life need not be the result of compromises with contingencies that are allowed unjustifiably to influence life prospects. Freedom from these influences does not mean that their choices about their lives are arbitrary or capricious. Rather conditions of justice leave members of the WOS free to form, pursue, and revise their plans in response to what they see as good reasons having what they take to be a legitimate bearing on their choices. Now let's grant Rawls that the WOS would encourage its members to think of themselves as free and to think that full autonomy is the most appropriate realization of their natural freedom. Then members of the WOS will value the freedom they have to plan their lives in a just society. Thinking of themselves as naturally free in this way, and wanting to exercise their freedom, they will want to follow their practical reason where it leads them within the constraints of justice.[11]

But where does practical reason lead?

10. Rawls, *Restatement*, p. 187.

11. See Rawls's instructive and suggestive remark about how Kant thinks we act "under the idea of freedom" at *Lectures in the History of Moral Philosophy*, p. 299: "in acting under the idea of freedom, we must regard our reason as free and guided by its own principles. The same must hold for what we count as reasons and their relative weight."

What Rawls calls "the burdens of judgment" are most often discussed as sources of reasonable disagreement that members of the WOS must acknowledge if they are themselves to be reasonable. Rawls speaks of them prominently in that connection in his later work (*PL*, pp. 54–58), and the most sophisticated commentary has generally followed suit.[12] But the burdens of judgment were originally introduced in *TJ* simply to explain the fact of pluralism to Rawls's readers, and not as an explanation of pluralism that members of the WOS themselves must accept (*TJ*, p. 127/110). Though I shall not pursue the matter here, the shift in Rawls's treatment is, I believe, a natural consequence of his political turn. But Rawls's original idea was that those who try to follow practical reason will reach different conclusions about what is good in life because, as subject to the burdens, their conscientious exercise of free reason will lead them to different conclusions. If this is correct, then Rawls does not simply assume that the WOS would be pluralistic, let alone extrapolate to this conclusion from the observed pluralism of extant liberal democracies. Rather, he thinks the pluralism of the WOS has its origins in the view of themselves that members of that society would be encouraged to adopt by their public conception of justice.

Pluralism about what is good in life poses a serious problem for the congruence arguments of *TJ*. Crudely put, it does so because the reasons and arguments members of the WOS have for maintaining their sense of justice is part of their good all depend upon their thinking of themselves and others as in various ways free, and in valuing activity—their own activity and that of others—that is free in those ways. But people who think of themselves as free to follow practical reason where it leads may not all converge on the same conception of themselves, and may not all value all the relevant kinds of freedom. Indeed, it is unrealistic to suppose that all of them will converge on the conception of their nature and their freedom that some of the congruence arguments require.

To see this, let's return to Joan and her reasons for maintaining her sense of justice, as laid out in Chapter V. According to the arguments reviewed there, she has the four desires referred to by C_4a, C_4b, C_4c, and C_4d. She wants to express her nature as a free and equal rational being. She wants to avoid the costs of hypocrisy and deception. She wants to protect persons and institutions she cares about. And she wants to take part in forms of association that call forth talents. I am going to postpone the problems pluralism poses for the argument from C_4a until the next section. Here I will concentrate on the problems it poses for the arguments from C_4b, C_4c, and C_4d. Pluralism does *not* undermine the claim that members of the WOS would have the desires to which those three conclusions refer, but it *does* poses difficulties for Rawls's claim that, in the special conditions of the WOS, those desires provide decisive reasons to be just.

12. See, for example, Thomas Nagel, "Moral Conflict and Political Legitimacy," *Philosophy and Public Affairs* 16 (1987): pp. 215–40, pp. 234–35.

Rawls argues that the second of these desires provides Joan a reason to be just because the only way she can be sure of satisfying that desire is to maintain her sense of justice as supremely regulative. But we saw in §V.5 that the force of this reason depends upon Joan's not attaching especially high value to wealth above her fair share. I suggested that if she does not attach especially high value to wealth above her share, it is at least in part because she knows that the size of her share depends upon her free choice of occupation. The fact that her occupation, and hence her share, were chosen freely under just conditions is salient for her when she asks whether she would rather have more. Even in a pluralistic society, it may be that everyone would still regard this kind of freedom as significant, and that everyone would regard it as having enough value to play the role that it must if the argument from C_4b is to succeed. I am prepared to grant that they would.

But now consider the argument from C_4d, the desire to take part in forms of association that call forth talents. We saw in §IV.2 that the presence of that desire depends upon the qualified form of the Companion Effect to the Aristotelian Principle:

> (4.5') "When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept" (*TJ*, p. 523/459).

The conditions under which (4.5') says people "are disposed to enjoy the perfections of others" are the conditions of a social union. *What* people enjoy about the perfections of others in a social union is that others develop parts of the nature that all of us share.

Now recall why the desire to take part in such associations provides Joan a reason to maintain her sense of justice as supremely regulative. It provides her a reason because she can participate fully in a social union of social unions only if she treats her sense of justice that way and because—as we saw in §V.3—the goods of a social union are available "to a preeminent degree" in a social union of social unions (*TJ*, p. 571/500). This, as we saw, is because in a social union of social unions, our latent powers are brought more fully to fruition than in smaller social unions and the diversity of activity is richer.

As I suggested in §V.5, some members of a pluralistic society may endorse conceptions of the good according to which some of the activities of others—even activities consistent with the principles of justice—are wrong or offensive. They may find others' ways of life frivolous or banal. They may consider their sexual practices immoral. They find their religious practices cultic or superstitious. If they are still to enjoy the richness of human activity available in a social union of social unions and to value their role in eliciting that diversity, it cannot simply be because they see all the ways of life that others' choose as valuable expressions of human nature. Rather, what they value about at least some others' ways of life must be that those lives were freely chosen by those who live them. In some cases, they must think, it is the choosing and not

the substance of the choice that manifests our common nature. This is supported, I believe, by Rawls's implication that members of the WOS see themselves and one another "as primarily moral persons with the equal right to choose their way of life."[13] And so what members of the WOS must value about their own participation in a social union of social unions is that their participation makes free choice among diverse lives possible. As I suggested in §V.5, the fact that they value others' free choices must be one of the reasons Rawls accepts (4.5').

The problem with this line of thought is that in a WOS, some people may have conceptions of the good according to which the exercise of freedom to choose a life that is trivial, immoral, or superstitious is not itself of value. Or at least, adherents of some conceptions will not regard the freedom to choose such lives as valuable enough for them to care about or take satisfaction in their own full participation in a social form that makes that freedom available. (4.5') seems not to be true of them. If there are such people then—even if they have the desire referred to by C_4d—it will not be true that they can best satisfy that desire by participating in a social union of social unions. It may be that those people can best satisfy that desire by participating in smaller social unions. In that case, the desire to take part in social forms that elicit talents will not provide them a reason to maintain their sense of justice as supremely regulative. They may have some other reason to be just, but the argument from C_4d will fail.

My claim that Rawls became dissatisfied with the argument from C_4d is confirmed, I believe, by fact that that argument does not appear in the places in his later work where we would normally expect to find it: in connection with the argument that participation in a just political society is experienced as a good.

The *locus classicus* of that argument in Rawls's later work is "Priority of Right and Ideas of the Good." In that essay, Rawls says that a WOS has a shared final end the attainment of which is highly valued, namely "establishing and conducting reasonably just democratic institutions" (*PL*, p. 204). Rawls does not say as clearly as we would like that the activities by which this achievement is gained are appreciated as good in themselves, but I believe this is what he thinks. If so then, taken together, these remarks imply that the WOS has the defining features of a social union: "shared final ends and common activities

13. Rawls is speaking of parties in the OP when he says at *TJ*, p. 563/493 that "their fundamental interest in liberty and in the means to make fair use of it is the expression of their seeing themselves as primarily moral persons with the equal right to choose their way of life." Since parties in the OP represent citizens in the WOS and act on their behalf, I take it Rawls thinks that interest stems from the way members of the WOS see themselves and one another: "as primarily moral persons with the equal right to choose their way of life." If this is the way members of the WOS see each other, then it stands to reason that part of what they would value about their participation in a social union of social unions is, as I have said, that it makes free choice possible.

valued for themselves" (*TJ*, p. 525/460). Moreover, Rawls says that in a WOS, citizens are secure in their ability to exercise their moral powers and experience their exercise as a good (see *PL*, pp. 202–3). Thus even after making his political turn, Rawls argued that the WOS is a form of life (i) with aims all accept, (ii) in which the activities of all have an agreed-upon place, and (iii) in which each is secure in the enjoyment of the exercise of his own powers.

The discussion of a social union of social unions in *TJ* might lead us to expect that at this point, Rawls would appeal to (4.5'), the qualified version of the Companion Effect. And we might expect him to conclude that the good of political society includes enjoyment of the "fruition of our latent powers" in "the perfections of others" and of "the greater richness and diversity of collective activity." And we might expect him to move from that conclusion to the conclusion that these goods of political society provide members of the WOS reason to participate in political society by maintaining their sense of justice. But that is not what Rawls does. He says that the activity of establishing and maintaining just institutions is a very great *political* good. He implies that members of the WOS will normally maintain *that* good as part of their plans of life. But *the good of pluralism itself* drops out. The fact that Rawls does not reproduce the argument from C_4d in his later discussion of the good of political society suggests the dissatisfaction with the argument from C_4d that I have alleged—a dissatisfaction that ultimately stems, I believe, from a loss of confidence in (4.5').[14]

What of the argument from C_4c? The desire referred to by C_4c—the desire for friendship—provides Joan a reason to maintain her desire to act from the principles because maintaining ties of friendship requires that we be just to our friends and because these ties are supposed to "extend rather widely, and include ties to institutional forms" (*TJ*, p. 571/500). If what I have said about

14. In "Social Unity and the Primary Goods," Rawls himself gives example of a WOS in which adherents of two different conceptions of the good each regard the others' conception with aversion and contempt; see *Collected Papers*, p. 381. The example did not lead him to give up the *Social Unions Argument* immediately, since the argument appears in the contemporaneous "Basic Liberties and their Priority" (*PL*, pp. 303–4, 371). By "Priority of Right and Ideas of the Good," written five years later, the argument has disappeared from his work.

I granted Rawls the argument from C_4b, which requires that people value their own free choice of occupation. Does my objection to the argument from C_4d imply that people value a freedom for themselves that they do not value for others? No. The argument from C_4b assumes that people are relatively satisfied with their fair share because they think that share reflects their own free choice among occupations. They think, "I am entitled to this much rather than more because I chose to spend my life as a teacher rather than a lawyer." This is consistent with their having made their choice based, in part, on their own views about what ways of life are worthwhile. It does not require them to think that if they had chosen a life that their conception of the good implies not worthwhile, that life would still have been valuable as well. The argument from C_4d, on the other hand, seems to require that they value and enjoy others' choice of lives that they themselves think are not worthwhile because even those lives are chosen.

the "social unions" argument is right, then extensive ties may not always be founded upon appreciation of diversity. But despite pluralism about the good, the disposition to reciprocity may insure that a kind of civic friendship obtains among members of the WOS. If so, then since the maintenance of basic institutions and of this kind of friendship requires justice, Joan will best satisfy the desire referred to C_4c by preserving her sense of justice as supremely regulative. The pluralism of the WOS may weaken the force of the reason established by the argument from C_4c, since ties of friendship in the WOS will be weaker than they would be if each member of the WOS valued the full range of human diversity. But I am supposing that the argument from C_4c itself remains a sound argument.

Thus, the arguments from C_4b and C_4c seem not to be undermined by the fact that some members of the WOS may not value complete freedom to choose within the bounds of justice. The desires to avoid the psychological costs of hypocrisy and to maintain ties of friendship provide Joan reasons to maintain her sense of justice. The argument from C_4d, however, *is* undermined by pluralism. Rawls can continue to maintain that C_4d itself is true, and that members of the WOS—or persons generally—have the desire to which it refers. But he has to abandon the argument that that desire gives rise to a reason for maintaining the sense of justice—and he does abandon it, as we just saw.

Can Rawls still rely on the *Argument from Love and Justice* to show that Joan's remaining reasons tell decisively in favor of maintaining her sense of justice? As we have seen, a fuller appreciation of pluralism implies that the reasons Joan has to maintain her friendships are weaker than they seemed when Rawls thought he could rely upon the argument from C_4d. Civic friendship, now understood as a relation among members of the WOS founded on the giving and receiving of justice, may seem a weaker bond than one founded on appreciation for the ways in which members develop the excellences of their common nature. The relative weakness of the bond raises a serious worry about the *Argument from Love and Justice*.

We saw in §VI.5 that the *Argument from Love and Justice* depends on a *Balance Conditional* that says:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of committing to her loves—including the wide-ranging attachments referred to by C_4c and C_4d—in the WOS.

The difficulty with the argument from C_4d means that that *Balance Conditional* would have to be altered to remove the reference to a social union of social unions, so that it says:

> If Joan would judge that her balance of reasons would tilt in favor of answering love with love in the world as it is, then she would judge that her balance of reasons tilts in favor of committing to her loves—including the wide-ranging attachments referred to by C_4c—in the WOS.

But once the *Balance Conditional* is altered in this way, it is questionable whether it is true. For the truth of the original *Balance Conditional* depended upon the transformative character of certain relationships. Those relationships, once entered into, transform one's structure of motives. Rawls may have been correct to assume that someone taking full part in a social union of social unions would find herself transformed in this way by her appreciation of the diversity of human excellence. But people who do not appreciate all the diversity of a liberal society will not be transformed in this way. If they are to be transformed, it will be because they appreciate those who treat them with justice. It is surely questionable whether this will be enough to sustain the *Balance Conditional* and Rawls's conclusion. In the world as it is, Joan's reasons surely tilt in favor of maintaining her relationships with those she loves, such as her children. Suppose that in the WOS, she could pass on covert, unjust gains to her children but would then know that she is not really living on terms of civic friendship with everyone who is just to her, including those whose choices about life she regards as fundamentally wrong-headed. Is it now so clear that *that* is the choice she would regret?

Thus once civic friendship is seen to be a weaker bond than it seemed, it is at least questionable whether civic friendship is as similar to love in the world as it is as the truth of the *Balance Conditional* requires it to be. If it is not, and if the altered *Balance Conditional* cannot be established, then Joan's remaining reasons to be just—the reasons connected with C_4b and C_4c—cannot be shown to be decisive.

This problem with the revised *Argument from Love and Justice* reflects the inherent difficulty of the original version of the argument, the difficulty that Rawls himself conceded. Even in its strongest form, the *Argument from Love and Justice* depends upon the simplifying assumption that participation in civic friendship is "an all or nothing decision" (*TJ*, p. 574/503). The difficulty with that argument was supposed to be overcome by the *Kantian Congruence Argument*, which is supposed to show that the reason connected with C_4a—the reason provided by our desire to express our nature—is decisive. I now want to look at the implications of pluralism for that argument.

§VIII.4: The Failure of Kantian Congruence

When I considered the arguments from C_4b, C_4c, and C_4d, I granted that members of the WOS have the desires to which they refer, and asked whether those desires provide reasons to maintain the sense of justice that can be shown decisive. I shall also grant that members of the WOS have the desire referred to by C_4a, the desire to express their nature as free, equal, and rational. In this case, however, I shall revisit the concession later. For the desire to express our nature differs in an important respect from the desires referred to by C_4b, C_4c, and C_4d. The presence of the latter desires depends upon the Aristotelian

Principle, the Companion Effect, and the laws of moral psychology. But the presence of the desires themselves—as opposed to their scope—does not depend upon special features of the WOS or of a liberal democratic society. By contrast, as I argued in §IV.2, the desire to express our nature as free, equal, and rational *does* depend upon the institutions and culture of liberal democracy. In Chapter IX, I shall suggest that this dependence raises an important question: is the nature members of the WOS want to express their nature as *persons* or is their desire more accurately described as a desire to express their nature as *citizens*? If the latter, then C_4a would have to be revised accordingly. For now, however, I shall grant that C_4a is true, and that members of the WOS all have the desire to express their nature as persons.

We saw in Chapter VII that a crucial step in the *Kantian Congruence Argument* is *TJ's Nash Claim*:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

and that the argument moves from C_4a to the *Nash Claim* via:

> (5.7') Joan's desire to express her nature moves her to treat her sense of justice as supremely regulative of her other desires.

and

> (5.9') "The desire to [treat her sense of justice as supremely regulative] and the desire to express [her] nature as free moral persons turn out to specify what is practically speaking the same desire."

And we saw that Rawls thinks he can rely on (5.7') and (5.9') because he thinks Joan, who grasps justice as fairness by the publicity condition, will see that:

> (5.5') The only, and hence the best, way for Joan to satisfy the desire asserted in C_4a is to treat her sense of justice as supremely regulative of her other desires.

We saw in §VII.2 that to accept (5.5'), Joan must accept and value a particular conception of her freedom. For the *Kantian Congruence Argument* to succeed, she must value *thin autonomy*. What I said about the *Argument from Love and Justice* might suggest a problem with the *Kantian Congruence Argument*: even in a society in which everyone holds a *free-and-equal self-conception* and wants to live as a free and equal rational being, ethical pluralism opens the possibility that there are people who deny that living *freely* requires the realization of thin autonomy. They will want to live freely, but will have a different conception or ideal of freedom that they want their lives to realize. And so even in a society in which C_4a is true, such as the WOS, there may be people who deny (5.5'). If Joan, who is the crucial case for congruence, is one of these people, then (5.7')

and (5.9') will not be true, and the *Kantian Congruence Argument* will fail. Since the *Kantian Congruence Argument* was needed to remedy an inherent weakness in the *Argument from Love and Justice*, Rawls's treatment of congruence would then need to be rethought.

But *why* might Joan deny that realizing thin autonomy belongs to her good? And *why*, exactly, does her denial undermine the argument? Is it just because she denies (5.5'), which implies that the best way to live freely is to live autonomously, so that (5.7') is untrue? Or does Joan's denial of (5.5') reflect a more profound rejection of the values and claims on which the *Kantian Congruence Argument* draws? To answer these questions, it is useful to recall the details of the argument.

Recall that (5.5') depends upon:

(5.2) The desire to express our nature is a desire to act from principles that would be chosen in the OP.

and that (5.2), in turn, depends upon

(1.9) The OP is a choice situation in which our nature is the decisive determining element.

We saw in §VII.4 that Rawls would defend (1.9) and (5.2) by a number of arguments, including arguments that connect the finality condition with an interest we are naturally said to take in the unity of our practical reason. Taken together, the *Kantian Congruence Argument* and the arguments that support its premises are supposed to establish that taking the desire to act from the principles as supremely regulative is good for members of the WOS when they follow the thin theory, and good for them because of the kind of beings they are: beings whose lives and powers of practical reasoning need to be unified in various ways. If they know themselves and their rational nature, then—according to the *Kantian Congruence Argument*—they will know how their practical reason is to be unified, and they will affirm that taking the desire to act from the principles as supremely regulative belongs to their good. In that case, (5.9') will be true and the argument will succeed.

The real problem with the *Kantian Congruence Argument* is that under conditions of pluralism, Joan—who, we are supposing, thinks of herself as free and rational—may endorse a conception of her nature that is incompatible with (1.9) and (5.2). The incompatibility of her conception of her nature with (1.9) and (5.2) is, I am supposing, what could lead her to reject (5.5'), so that (5.7') and (5.9') would be untrue. *That* is what undermines the *Kantian Congruence Argument*.

I want to explore the possibility that Joan endorses views of her nature that are incompatible with (1.9) and (5.2) *because* she denies we have a rational interest to which the finality condition answers. I took the finality condition to impose two requirements. *Ultimacy* requires that principles chosen in the OP function as the final arbiters of competing claims; their verdict is not to be checked against further principles. *Perpetuity* requires that the principles, once

adopted, hold "once and for all"[15] (*TJ*, p. 176/153). I shall briefly explore views that deny the connection of each with our nature.

Suppose Joan thinks that human beings are created to love and serve God. I am supposing that Joan has interests in the *Unity of Character* and the *Consistency of the Right and the Good* that I discussed in §VII.6. But Joan takes seriously the Pauline lament "For what I am doing, I do not understand. For what I will to do, that I do not practice; but what I hate, that I do."[16] Paul thinks that what he describes is typical of the human condition. The experience he describes as typical seems to be that of acting contrary to the law, despite his desire to obey it, because of desires or impulses that he himself regards as irrational. Paul thus thinks that each person is divided within herself. The impulses to sin that Paul discusses pose, Joan thinks, the biggest threat to her unity of self. Because of the way Joan conceives of her nature, she thinks the right way to achieve the *Unity of Character* and the *Consistency of the Right and the Good* is to regulate her life by divine commands, and to make service to the will of God her ultimate end—a position she thinks is caricatured by Rawls's critique of dominant ends. Other principles from which she acts, such as those chosen in the OP, are subordinate to divine commands rather than ultimate. She must treat them as such if she is to achieve real unity of self.

Rawls says that he has set up the OP so that our interest in the unity of the self helps to determine choice there. From Joan's point of view, however, he has not. So Joan rejects one of the arguments that supports (1.9), as well as (1.9) itself. This undercuts Joan's support for (5.2). Moreover, Joan's view of human nature may carry with it a different conception of freedom than the one the *Kantian Congruence Argument* requires. To be free, Joan may think, requires overcoming sinful impulses and willing the good. This kind of freedom, Joan may think, requires submission to divine authority. If this is what she thinks, then Joan will believe that she can express her nature as a free being only insofar as she follows divine commands and treats them as ultimate. So she denies that the desire to express our nature is the desire to act from principles that would be chosen in the OP. She therefore rejects (5.2) and (5.5'), and (5.7') and (5.9') will be untrue.

I do not believe that the possibility of Joan's having the religious views I just ascribed to her is what convinced Rawls that the *Kantian Congruence Argument* failed and that his treatment of congruence needed to be rethought. Even so, I think the case is of considerable interest. It shows how the *ultimacy* condition on principles can fail to answer to the precise interest in unity of the self that some members of a pluralistic society might take themselves to have. Furthermore,

15. We saw in §VII.6 that Rawls sometimes treats the two conditions as if they were imposed separately, and equates finality with *ultimacy*. Even if we follow the passages in which he does so, *perpetuity* will still be an element of the OP that reflects our nature. So long as the demands of the two conditions are not conflated, there is no difficulty in speaking as if both follow from finality.

16. *Romans* 7, 15.

considering the range of possible responses to the case helps us to understand some of the salient features of political liberalism. I shall return to these points in Chapter IX. First, I want to look at the kind of case I do think may have led Rawls to become dissatisfied with the *Kantian Congruence Argument*.

Rawls assumes that members of the WOS have a conception of themselves as living lives that are extended in time over a normal span. Rawls implies as much, using the term "ideal" instead of "conception," in an important remark in "The Independence of Moral Theory":

> the ideal is that of persons who accept responsibility for their fundamental interests over the span of a life and who seek to satisfy them in ways that can be mutually acknowledged by others.[17]

Members of the WOS are therefore thought of as having a rational interest in unifying their plans over a complete life. The *perpetuity* condition insures that that interest helps to determine choice in the OP. But now suppose that Joan does not think of herself as persisting throughout the course of a bodily life. Suppose, instead, that she thinks of herself as a Parfitian self.[18] Suppose, that is, that she thinks personal identity is not a deep fact about her that holds over and above bodily and psychological continuities and connections. Rather, her identity over time just consists in these connections, which may hold to reduced degrees. Then the interest in imposing lifelong unity on her plan is not an interest she has in virtue of having a rational nature. Rather, insofar as she is rational, she may have some different and contrary interest: perhaps she is interested in living moment-to-moment, in imposing unity on segments of her life, or in unifying some segment of "her" plan and with that of "someone else." The way the OP is structured, the interest she has as a Parfitian self does not help to determine the choice of principles. If Joan thinks of herself in the Parfitian way, then, she would reject the argument for (1.9) and (1.9) itself.

What will she make of the other critical steps in the *Kantian Congruence Argument*—(5.2), (5.5'), (5.7'), and (5.9')? Even if we suppose that beings with a Parfitian self-conception think of themselves as free, equal, and rational and want to express their nature as such, it is hard to know what the object of that desire would be. It is hard to know, for example, what conception of freedom is most appropriate for beings who think of themselves in this way. But without (1.9), there is no reason to think that a being with a Parfitian self-conception, as I am now supposing Joan to have, would accept (5.2) and (5.5'). Hence there is no reason to think that (5.7') and (5.9'), which asserts that Joan's desire to express her nature moves her to treat her sense of justice as supremely regulative, are true. Indeed, if (5.7') and (5.9') ultimately imply that that desire

17. Rawls, "Independence of Moral Theory," *Collected Papers*, p. 299.

18. See Derek Parfit, "Later Selves and Moral Principles," in *Philosophy and Personal Relations: An Anglo-French Study* (London: Routledge & Kegan Paul, 1973), ed. Alan Montefiore, pp. 137–69.

moves her to maintain her sense of justice as regulative over a complete life, then there is some reason to doubt it.

I suggested that this may be the kind of case that convinced Rawls the *Kantian Congruence Argument* failed because I think it is the kind of case that is suggested by the article of Samuel Scheffler's that Rawls credits with leading him to rethink his early treatment of stability (*PL*, pp. xxxiv–xxxv). In this case as in the first, the *Kantian Congruence Argument* fails and it founders on the same problem. Joan may grant that if a Kantian conception of human nature were correct, then (1.9) would be true. And she may grant that (1.9) will *seem* to be true to someone who thinks of herself in this way. But the *Kantian Congruence Argument* requires something stronger than these concessions. The argument requires that Joan grant (1.9) itself. That is the claim that I am now imagining that Joan rejects. I am imagining that she rejects it because, thinking of herself as free to follow her practical reason, she arrives at a very different view of her nature and of how her life is to be unified.

The demands of justice are the demands or requirements of living up to a view that members of the WOS have of themselves. They are the demands of the *free-and-equal self-conception* initially expressed by (1.1) and specified by the ideal of full autonomy. Views of the self are not implicit in human rationality, waiting to be discovered by philosophical reflection. They are shaped by the basic social institutions under which we live. Part of what a theory of justice must do is identify the view of ourselves that just institutions should encourage, and connect that self-conception with principles of justice. Part of what a just society must do is educate its members in that view of themselves, since stability depends upon their wanting to live up to it. But in the modern world, just societies encourage their members to think of themselves as free to choose among conceptions of the good and among the conceptions of persons that go along with them. This opens the question of whether there is any conception the person that both leads to appropriate principles of justice and is such that everyone in a just society wants to realize it. Rawls came to believe that *TJ* had not adequately answered that question.

§VIII.5: The Great Unraveling

Let me now sum up the problems Rawls found in *TJ*'s treatment of stability.

The Rawls of *TJ* thought that the inherent stability of the WOS depended upon the congruence between the right and the good. In Chapter III, I argued that Rawls posed the question of congruence in two ways in *TJ*. The first, understood in light of the *Deweys*, was:

> Is it rational for Joan to maintain her sense of justice on the basis of her desires for objects valued according to the full theory, including the objects of her ideal-dependent desires?

I have said that Rawls argued for:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

or, more weakly, for:

C_3*: All members of a WOS want to live up to the ideal of full autonomy.

If members of the WOS all want to live up to those ideals and if their desires are strong enough, then the first congruence question would be "yes" and Rawls would have established what I called the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

The *Congruence Conclusion* describes a state of affairs that is in equilibrium. Since the ideal-dependent desires referred to C_3 and C_3* would be enduring, each person's commitment to justice would be stable and members of the WOS would regulate their plans by the principles over the course of life. The equilibrium would be stable.

Rawls's argument for C_3* depends upon members of the WOS all accepting crucial claims in the *Kantian Congruence Argument*. His argument for C_3 depends upon their accepting those claims and on the truth of (4.5'). As we have seen, C_3 and C_3* entail that members of the WOS would converge on a partially comprehensive doctrine on the basis of which they would affirm that maintaining their desire to treat the principles as supremely regulative belongs to their good. The institutions that implement and publicize justice as fairness encourage this enduring convergence. That is one of the ways they generate their own support and it is why the stability that results is inherent stability. So the Rawls of *TJ* thought that the institutions of the WOS bring about the truth of C_3.

As I have tried to show, Rawls came to realize that those institutions also encourage pluralism. Realizing this, he came to think it unrealistic that members of the WOS would accept the claims on which the *Kantian Congruence Argument* depended, and he came to think (4.5') was unrealistic as well. The upshot was that not everyone in the WOS was likely to aspire to the ethical ideals to which C_3 refers, for these ideals may conflict with other ethical ideals and views of themselves that citizens endorse under conditions of pluralism. So the institutions that are supposed to bring about the truth of C_3 also bring it about that C_3 is likely to be false. The argument from C_3—or C_3*—to C_C fails.

This inconsistency, and the failure of the argument for C_C, would themselves be severely damaging to Rawls's account of stability. C_C refers to judgments that members of the WOS make from the viewpoint of full deliberative rationality. That is the viewpoint they adopt, or try to adopt, when they make

their plans. I have conjectured that in later work, Rawls—with some justification—read *TJ*'s treatment of stability as if it relied on C_3 or C_3*. If I am right, then the Rawls of *TJ* and the original *Deweys* thought that members of the WOS would take account of ideal-dependent desires in drawing up their plans of life. Their commitment to justice in daily life—and hence the stability of justice as fairness—depended upon the presence and strength of these desires.

The objects of these desires are, of course, ideals that belong to justice as fairness. They are part of its theoretical apparatus. They are part of the sense of justice it encourages and the full theory of the good accounts for their value. Because they are part of justice as fairness, and because they are sufficient to stabilize it, Rawls later said *TJ* took justice as fairness to be stable because "the political conception of justice is maintained as in itself sufficient to express values that normally outweigh . . . whatever values might oppose them." When members of the WOS all have the relevant ideal-dependent desires, stability is a straightforward matter—that is why Rawls later said that *TJ* treats of the "simplest case" of stability.[19]

The inherent stability of justice as fairness would obviously be threatened if the institutions that are supposed to encourage such desires would in fact undercut them. Reflecting on *how* institutions undercut those desires shows the extent of the damage. The question of congruence arises because of the possibility that members of the WOS—who have a sense of justice and are therefore moved by ideal-dependent desires—would be tempted to treat their sense of justice as one more desire which can be traded off against desires for other things they want.

This concern is not answered by asserting that members of the WOS have an effective sense of justice or sufficiently strong ideal-dependent desires. That, as we saw, is why Rawls thought the first form of the congruence question "has an obvious answer" (*TJ*, p. 569/498); it has an obvious answer despite the great interest of the arguments Rawls offers in defense of the claim that they have such desires. Whether members of the WOS would give in to temptation, or preserve their sense of justice as supremely regulative, depends upon what else they would want and what other ends they have.

Rawls tried to answer this concern by arguing that just institutions would encourage the four desires I discussed in Chapter 4, the desires referred to by:

C_4a: All members of the WOS think of themselves, at least implicitly, as naturally free, equal, and rational persons, and want to express their nature as such.

C_4b: All members of the WOS want to avoid the psychological costs of hypocrisy and deception.

C_4c: All members of the WOS want ties of friendship.

19. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 414, note 33.

C_4d: All members of the WOS want to participate in forms of social life that call forth their own and others' talents.

and by moving from C_4a, C_4b, C_4c, and C_4d to *TJ's Nash Claim*:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

We have seen how Rawls moves from *TJ's Nash Claim*, via a solution to the *mutual assurance problem*, to:

C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

From C_6, Rawls can infer the *Congruence Conclusion*. Since the desires referred to by C_4a, C_4b, C_4c, and C_4d, like the ideal-dependent desires, would be enduring, each person's commitment to justice would be similarly enduring, the equilibrium described by the *Congruence Conclusion* would be stable, and the WOS would be stably just.

The move from C_4a, C_4b, C_4c, and C_4d to C_N, C_6, and the *Congruence Conclusion* would have enabled Rawls to show inherent stability by defending an affirmative answer to the second, and more interesting, form of the congruence question:

> Is it rational for Joan to maintain her sense of justice on the basis of the desires referred to by C_4a, C_4b, C_4c, and C_4d?

The arguments that were supposed to establish C_N and C_6 were the *Argument from Love and Justice* and the *Kantian Congruence Argument*. But as we have now seen, Rawls came to think C_3 is unrealistic *because* he came to think that by encouraging pluralism, the institutions of the WOS would make it less likely that those who live under them would accept the claims they would have to accept for those arguments to succeed. Thus, "the account of stability in part III of *Theory* is not consistent with the view as a whole" (*PL*, pp. xvii–xviii) because *TJ* "fails to allow for the condition of pluralism to which its own principles lead"[20] when they are implemented.

Because of the inconsistency in justice as fairness, the Rawls of *TJ* could not show that a plan of life regulated by principles of justice is each person's "best reply to the similar plans of his associates" (*TJ*, p. 568/497). This is *TJ's Nash Claim*. Without that claim, the Rawls of *TJ* could not show that a just society would be in equilibrium. Nor could he show that no one in the WOS

20. Rawls, *Restatement*, p. 187.

would have sufficient reason to defect from the agreement that would be reached in the original position. Public knowledge of the latter was crucial to solving the *mutual assurance problem*. Since showing the inherent stability of justice as fairness required showing that such an equilibrium would be stable, and showing it by solving the *mutual assurance problem*, *TJ*'s argument for the inherent stability of justice as fairness failed.

The unlikelihood of C_3 or C_3^* raised a question that Rawls did not confront in *TJ*:

> If some members of the WOS do not have the ideal-dependent desires implied by C_3 or C_3^*, is it rational for them to maintain their sense of justice—when others maintain theirs—on the basis of the various comprehensive views of the good they *do* hold?

If the arguments from C_4a, C_4b, C_4c, and C_4d to C_N, C_6, and the *Congruence Conclusion* had been successful, Rawls could have answered this question by answering the congruence question in its second form. But they were not successful, and Rawls made the changes between *TJ* and *PL* to show equilibrium on other grounds. After he made those changes, Rawls continued to believe that stability—understood now as "stability for the right reasons"—depended upon the presence of enduring ideal-dependent desires and that those desires would be encouraged by just institutions. But the Rawls of *PL* thought that the objects of those desires were ideals that were "political not ethical."

As we shall see, this means that in *PL*, the case for stability had to appeal to variants of C_3 and C_N, rather than to C_3 and C_N themselves. The Rawls of *PL* still relies on a *Nash Claim* to establish inherent stability, but the *Nash Claim* on which he relies, and the solution to the *mutual assurance problem* that he develops, are very different from those in *TJ*.

We can gain a somewhat different view of how the inconsistency in justice as fairness arises by thinking again about what inherent stability requires. I initially said that it requires members of the WOS to maintain their sense of justice, rather than deciding case-by-case whether to accede to temptations that arise from the self-interested point of view. The examples of such temptations that I gave then were the temptations to cheat on one's taxes or to desert one's post. I used these examples because I think these are the kinds of cases the Rawls of *TJ* had in mind.

The focus on such garden-variety moral failures—and the way allegiance to moral ideals can eliminate them—suggests that in his treatment of stability, at least, the Rawls of *TJ* and the original *Deweys* was doing what Bernard Williams accused him of: neglecting real politics and treating political philosophy as "applied moral philosophy."[21] But it is important not to interpret "the self-interested point of view" and its temptations too narrowly. When I introduced C_N and C_6, I cautioned that Rawls would eventually give the idea of that

21. Williams, *In the Beginning Was the Deed*, p. 77.

point of view considerable refinement. As he came more deeply to appreciate the fact of pluralism, he came to realize that what really need to be removed are *any* temptations to act against the sense of justice that arise from within any other point of view than "the point of view of justice" (*TJ*, p. 568/497) as defined by justice as fairness.

These other points of view include that of people who want to cheat on their taxes to have more money for themselves, and that of soldiers who are tempted to desert their posts so as to live another day. They include that of parents who are tempted to cheat on their taxes so as to accumulate money for their children, and soldiers who are tempted to desert so that they can go home to care for aging parents. They also include the points of view afforded by various identities that might lead people sincerely to make claims that are contrary to Rawls's principles. Some members of historically oppressed racial or ethnic groups may think that fair equality of opportunity or the fair value of the political liberties are not enough to make up for the legacy of their suffering. Some members of the WOS may sincerely think that they can be faithful to their religion only if they press for some restrictions on the liberty of those of other faiths. If justice as fairness is to be inherently stable, then members of the WOS must willingly put these claims aside because they see that the balance of their reasons favors maintaining and acting from the principles of justice. According to *TJ* and the original *Deweys*, their balances tip in favor of Rawls's principles because the institutions of a WOS encourage them to value the ethical ideals of justice as fairness above their ethnic, racial, or religious identities. But in a society whose institutions also encourage ethical pluralism, this seems unlikely.[22]

The unraveling of Rawls's early treatment of congruence had profound effects on justice as fairness and on his hopes for political philosophy. Two of these affects stem from Rawls's loss of confidence in (4.5'), the qualified version of the Aristotelian Principle's Companion Effect on which Rawls relied in the *Social Unions Argument*.

The loss of confidence in (4.5') affected Rawls's views about the quality of public life in the WOS. In *TJ*, Rawls wrote that "for the purposes of justice [we are] to avoid any assessment of the relative value of one another's way of life" (*TJ*, p. 442/388). While it is possible to read this important remark as saying that we avoid such assessment "*merely* for the purposes of justice," the section on a social union of social unions discourages this reading. Instead it encourages an interpretation which soft-pedals the phrase "for the purposes of justice" by suggesting that members of the WOS enjoy the pluralism of their society and regard a great many different lives as valuable. Pluralism, it

22. Catherine Audard briefly considers the hypothesis that "Among the external sources [of Rawls's shift to political liberalism] …the new social movements of the 1980's brought to his attention questions of identity and culture that influenced rival conceptions of justice." See her *John Rawls* (Montreal: McGill-Queens University Press, 2007), p. 182. The text to which this note is attached is intended to show how that hypothesis might be defended.

suggests, is to be welcomed and celebrated. The disappearance of the argument from C_4d suggests that in the WOS of Rawls's later work, citizens may have very different attitudes toward pluralism. Some may still regard it as something to be celebrated. Others, however, may regard many of the lives that differ from their own as permissible but regrettable choices.

This led to a second important change, a change in what Rawls thought political philosophy can hope to accomplish. One of the tasks of political philosophy, Rawls says, is "reconciliation." One of the things we need political philosophy to reconcile us to, he says, is pluralism. Speaking of the fact of pluralism, Rawls says "this fact is not always easy to accept, and political philosophy may try to reconcile us to it by showing us the reason and indeed the political good and benefits of it."[23] This remark was written long after *TJ*, yet we can see how *TJ*'s discussion of the WOS as a social union of social unions promised one account of those benefits. It promised that we could be reconciled to a pluralistic world by seeing in the diversity of religious and ethical views a full realization of our nature.[24] Once Rawls recognized that some members of a pluralistic society might adopt views according to which pluralism itself is politically permissible but regrettable, the promise of achieving reconciliation this way had to be abandoned. Philosophy might reconcile us to pluralism by showing how pluralism necessitated the political goods of ecclesiastical disestablishment and freedom of conscience.[25] But that project would rely on very different arguments than the *Social Unions Argument*, arguments which are much easier to make and which therefore suggest a lowering of Rawls's ambitions for philosophy.

The inconsistency Rawls says he found between *TJ*'s account of stability and his view as a whole is deep. For the account of stability and the view as a whole allow inconsistent answers to a deep philosophical question about how members of the WOS must conceive of themselves if they are to think of themselves, and to act as, naturally free and equal rational agents. This inconsistency struck at the heart of the constructivist view Rawls was trying to develop. For in the original *Dewey Lectures*, Rawls described the task of identifying principles of justice for the WOS as that of identifying principles "for social cooperation among persons who conceive of themselves as free and equal moral persons."[26] It is the task of identifying principles "for social cooperation" among persons who have what I have called a *free-and-equal self-conception*. As we have seen, carrying off this task requires Rawls to specify that self-conception by fashioning a more precise conception of the person. Only with a more precise conception in hand will the conception of the person yield the appropriate principles.

23. Rawls, *Restatement*, pp. 3–4.

24. Rawls does gesture in this direction at *Restatement*, p. 76. But even on that page, he very quickly moves to the question of how we are to be reconciled, not to the fact of pluralism, but to the fact that "some are by nature better endowed than others."

25. See *PL* p. xxvi, text and note 11; *Restatement*, p. 34, note 25; *Lectures on the History of Moral Philosophy*, pp. 347–48.

26. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 309.

This connection between the conception of the person and the choice and content of the principles is what is distinctive about Kantian constructivism.[27] The fact that that conception of the person seems to articulate our considered view of ourselves better than the conception relied on by intuitionism and utilitarianism is supposed to be part of what tells in favor of constructivist views. If the conception of the person Rawls develops—the conception members of the WOS must accept if the right is to be congruent with the good—is inconsistent with views about themselves at which members of a pluralistic society might arrive, then Rawls's constructivism faces a very serious problem. Solving that problem required Rawls to re-present the conception of the person on which he relied as a political conception of the person, and to re-present justice as fairness as a political conception of justice.

In *PL*, Rawls avers "surprise" that recasting *TJ*'s account of stability "forces many other changes and calls for a family of ideas not needed before" (*PL*, p. xix). We have now begun to see why this is so. In Chapter IX, I shall try to indicate how the changes between *TJ* and *PL* respond to the problem Rawls found in the original presentation of justice as fairness. I want to close this chapter by noting how my account of those changes differs from others on offer.

§VIII.6: Brief Contrasts with Other Accounts

Rawls implies in "Political not Metaphysical" that if justice as fairness relied on the "comprehensive moral ideal[]" of autonomy, it would be "but another sectarian doctrine."[28] His writings from that essay onward show his recognition that in a pluralistic society, not everyone will endorse that ideal. Readers have long recognized that the changes between *TJ* and *PL* were prompted by his worry that he had relied on the ideal of autonomy in *TJ*.[29] What they have often failed to do is identify arguments in *TJ* that he thought were undermined by his reliance on that ideal. Someone who thought that a critical argument for the two principles—for example, what I called in Chapter I the "Pivotal Argument"—depended on an appeal to autonomy might think Rawls's increasing appreciation of pluralism led him to worry about the justification he had offered for the principles in *TJ*. But as we saw in §I.5, the Pivotal Argument does not require an appeal to autonomy. Furthermore, the arguments for the principles are offered in *TJ*, part I. The claim that Rawls's reliance on autonomy undermined some argument for the principles seems

27. See Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 304: "What distinguishes the Kantian form of constructivism is essentially this: it specifies a particular conception of the person in a reasonable procedure of construction, the outcome of which determines the content of the first principles of justice."

28. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 409.

29. See, for example, the review by Stephen Holmes cited at Chapter III, note 11.

not to take seriously his remark that the "serious problem internal to justice as fairness" is an inconsistency between "the view as a whole" and part III of *TJ* (*PL*, pp. xxvii–xxviii).

Readers who have taken the claim seriously have not, I believe, appreciated the full scope of the problem Rawls found in *TJ*. Samuel Freeman, for example, says that what was unrealistic about *TJ*'s account of stability was the *Kantian Congruence Argument*.[30] As I have tried to indicate, especially in §§V.2 and VII.3, I disagree with Freeman about exactly what the conclusion of the argument is, and about the role of the Aristotelian Principle in reaching it. I have tried to indicate in greater detail why Rawls thought the *Kantian Congruence Argument* failed, and to show that Rawls also thought that a crucial assumption of the *Argument from Love and Justice* was unrealistic as well.

In a very influential essay on Rawls, Burton Dreben expressed some skepticism that religious members of the WOS could endorse the ideal of full autonomy and he noted that some members of a pluralistic society like the WOS would reject Rawls's two principles of justice, though he did not explicitly connect the two points. Instead, Dreben argued for an inconsistency between the possibility that some members of the WOS would reject the two principles, and the account of stability in *TJ*, part III that presumes everyone accepts them. Dreben states this interpretation bluntly, saying:

> Now what Rawls began to see was that, under the very conditions that satisfy the principles of justice that he worked so hard to establish, reasonable and free and equal people will begin to differ, inevitably and properly so, on those very principles of justice. Hence, from his perspective, the theory of stability that he had set forth in the last third of the book contradicts the first two-thirds of the book.[31]

It may well be that members of a WOS *would* disagree about justice, as Dreben asserts. As Rawls modified his theory, he certainly left that possibility open (*PL*, p. 164). I suspect that if the conditions of the WOS did give rise to disagreements about justice, it would be because they gave rise to disagreement about the good and because, as Jeremy Waldron has emphasized, different conceptions of the good generally have different implications for justice.[32] But disagreement about the good can pose a more immediate problem, for it can undermine Rawls's congruence arguments in ways I have sketched.

Moreover, the possibility of disagreement about justice, cited here by Dreben, is not needed to explain the changes between *TJ* and *PL*. For the inconsistency Rawls found concerns (1.9), the claim that the OP is a choice situation in which our nature is the decisive determining element. We saw in

30. See Freeman, *Justice and the Social Contract*, p. 168.

31. Dreben, "On Rawls and Political Liberalism," p. 317.

32. See Jeremy Waldron, "Disagreements about Justice," *Pacific Philosophical Quarterly* 75 (1994): pp. 372–87.

§VII.8 that it is possible to reach the two principles while bypassing that claim, by requiring that principles of justice be acceptable at every social position.[33] What cannot be gotten without (1.9) is the *Kantian Congruence Argument*. It is the implications of pluralism for *that* argument and for the rest of his treatment of congruence—not the possibility of disagreements about justice—that led to the changes between *TJ* and *PL*. This conclusion derives some support from the fact that Rawls seems to have thought about disagreements about justice only *after* he began his political turn. Neither "Political not Metaphysical" nor the original version of "Idea of an Overlapping Consensus" raises the possibility of such disagreements.

My explanation of the changes between *TJ* and *PL* also differs significantly from the explanation offered by the *Public Basis View*.

In §I.9, I said that the *Public Basis View* offers too simplistic an account of the changes between *TJ* and *PL* because it works with too superficial an understanding of publicity and simply assumes that the WOS would be pluralistic. According to the *Public Basis View*, publicity ensures that members of the WOS are in a position to know how the principles of justice are derived. The fact of pluralism, which Rawls is said simply to have assumed, is then said to imply that some people might disagree with premises of the argument for the principles. The *Public Basis View*'s treatment of publicity is superficial because the *View* does not appeal to the educative effects of publicity that Rawls stresses in the original *Deweys*. Failing to note the educative effect of publicity, proponents of the *Public Basis View* fail to see that Rawls has an explanation for pluralism. For Rawls thought that the way members of liberal democratic societies are encouraged to think of themselves helps to explain the pluralism of those societies. To see the tension that Rawls found in justice as fairness, we need to see first how the self-conception encouraged by free societies gives rise to pluralism, and then how pluralism gives rise to views of the self and of freedom that are at odds with those the congruence arguments require.

I do not deny that Rawls modified claims which served as premises in the Pivotal Argument, such as (1.9) and—as we shall see—(1.1). But I do insist that those modifications were prompted by the difficulties Rawls found in his treatment of congruence. My explanation therefore takes Rawls at his word when he says that the changes between *TJ* and *PL* were motivated by problems he found in part III of *TJ*. The explanation offered here therefore avoids the textual difficulties that beset the *Public Basis View*.

The explanation also shows that Rawls has a more nuanced and less thoroughly cerebral understanding of stability than that suggested by the *Public Basis View*. If a WOS is to be stably just, the arguments offered for the

33. The arguments of §VII.9 showed that it is possible to move to Rawls's principles from (1.6)—the requirement that principles be "acceptable to us as free and equal persons"—while bypassing (1.9). But did they show the possibility of making that move regardless of how persons are conceived? They did if, as I suggested then, the alternative route from (1.6) to the principles is neutral among conceptions of the person; see Chapter VII, note 23.

principles of justice must be accepted and citizens' acceptance of those arguments is part of what secures stability. But conviction itself is complex, and stability depends upon a mix of intellective and affective components. It depends upon the widespread possession of a sense of justice, which is developed by a desire to emulate exemplars and which is, as we saw in §IV.2, connected with relationships of love and affection. A WOS is most stable when its members maintain the sense of justice as part of their good. As I have tried to show, Rawls thought the congruence arguments of *TJ* depend upon the ideal-dependent desire to conduct oneself as a certain kind of person. As we shall see when we look at how an overlapping consensus stabilizes, we will see that it, too, relies on a complex of stabilizing forces and that those forces prominently include ideal-dependent desires. In Chapter IX, I shall confirm my reading by showing how the most obvious changes between *TJ* and *PL* respond to the difficulties I have identified in Rawls's original treatment of stability.

IX

The Political Ideals of Justice as Fairness

We have now seen that the Rawls of *TJ* and the original *Deweys* offered an account of congruence, and hence of stability, that depended upon:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

Or at least on

C_3*: All members of a WOS want to live up to the ideal of full autonomy.

Because the ideals to which C_3 and C_3* refer are ethical ideals, Rawls eventually came to think that *TJ* treated justice as fairness as what he called in *PL* a "partially comprehensive doctrine." But at the time he published the original *Deweys*, he still thought that institutions well-ordered by justice as fairness could bring about convergence on those ideals. Indeed, he thought that bringing about this convergence was one of the ways that justice as fairness would stabilize itself. The stability of justice as fairness depends upon the congruence of the right and the good. We saw that if the members of the well-ordered society (WOS) have effective desires to live up to the ideals justice as fairness includes, and if they would judge that living up to them to belongs to their good, then Rawls could infer what I called the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

In Chapter VIII, we saw the problems with this line of thought. Just institutions encourage members of the WOS to think of themselves as free, and those who act under the idea of freedom may well reach conclusions about the good that are inconsistent with the partial conception of the good that C_3 and C_3* say they share. Both C_3 and C_3* are therefore too strong to be realistic, and cannot be used to support the *Congruence Conclusion*. The considerations that show C_3 and C_3* to be unrealistic also undercut the congruence arguments Rawls actually laid out in *TJ*, the arguments from C_4a, C_4b, C_4c, and C_4d, by showing that Rawls was unable to move from those four premises to what I called *TJ's Nash Claim*. *TJ's Nash Claim*, and public knowledge of it, were crucial to *TJ*'s argument for congruence and stability. Without the *Nash Claim*, Rawls cannot get to the *Congruence Conclusion* from the desires to express our nature, avoid hypocrisy, live as friends, and take part in social unions.

I said that the improbability of C_3 and C_3*, and the failure of the congruence arguments he offered in *TJ*, forced Rawls to confront a question he did not take up in his earlier work:

> If some members of the WOS do not have the ideal-dependent desires implied by C_3 or C_3*, is it rational for them to maintain their sense of justice on the basis of the various comprehensive views of the good they *do* hold?

On my reading, Rawls reframed justice as fairness as a political conception of justice so that he could defend an affirmative answer to this question. On the new account as on the original one, the stability of the WOS depends upon all members of the WOS wanting to live up to certain ideals or conceptions of themselves that are realized only when they maintain their desire to act from the demands of right. But Rawls tries to avoid the difficulties of the original account by weakening his claim about what ideals members of the WOS would want to live up to. Instead of arguing that stability depends upon their wanting to live up to *ethical* ideals, the Rawls of *PL* hoped to argue that stability depends only upon their wanting to live up to *political* ideals. And so in his new account of stability, he hoped to appeal, not to the unrealistic claim C_3, but to the weaker:

> C_3': All members of a WOS want to live up to the *political* ideals of conduct, friendship, and society included in justice as fairness.

It may have been unrealistic to suppose that members of the WOS would converge on the partial conception of the good to which C_3 and C_3* refer. It is not so unrealistic, he hoped, to claim they would converge on the political ideals referred to by C_3'. The challenge Rawls faced, then, was that of showing that C_3' is true and is strong enough to support a stability argument. In this chapter and the next, we shall see how he tried to meet this challenge.

I have said that the changes between *TJ* and *PL* result from Rawls's attempt to remedy the difficulties he found in his original treatment of stability. In the "Introduction," I listed the changes between *TJ* and *PL* that are in greatest need of explanation:

- The stability of a WOS is secured by an overlapping consensus of reasonable comprehensive doctrines.
- Justice as fairness is presented as a political conception of justice, founded on basic ideas drawn from democratic political culture.
- The conception of the person represented by the OP is said to be a political conception.
- The idea of public reasoning, which was hardly mentioned in *TJ*, is prominent in *PL*.
- The notion of political legitimacy, which received no explicit mention in *TJ*, assumes a very prominent role in *PL*.
- In *PL*, Rawls admits that the citizens of a WOS may endorse any of a number of liberal political conceptions of justice rather than justice as fairness alone.
- The attempt to show that justice as fairness would be inherently stable is replaced by an attempt to show that it would be stable "for the right reasons."

I also mentioned three other changes that are less obvious but very important: Rawls's description of the sense of justice and his argument that political society is a good undergo subtle but revealing changes, and the notion of congruence—so central to Rawls's treatment of stability in *TJ*—does very little work in *PL*. Some of these changes, such as the changed description of a sense of justice and the eclipse of congruence, are connected to the way Rawls set up the problem of stability in *PL*. I shall discuss these changes in §IX.2 and §IX.5. Other changes are connected to the argument he thought would solve the stability problem.

In §IX.1, I lay out what I take to be the main argument for stability in *PL*. As in *TJ* so in *PL*, a crucial step in the argument is a Nash claim—roughly, the claim that a plan regulated by the desire to be just is each person's best reply to the similar plans of everyone else. The centrality of a Nash claim to the treatment of stability is just what we would expect, as I said when I introduced *TJ's Nash Claim* in §II.3. The idea of reciprocity lies at the heart of Rawls's account of justice. The WOS is a scheme of social cooperation that is organized on fair terms. Fair terms of cooperation are "terms that each participant may reasonably accept, provided that everyone else likewise accepts them" (*PL*, p. 16). The disposition to act from those terms is therefore a disposition to reciprocate or to respond in kind. If the cooperative scheme is to be stably just, each participant must be able to see, on reflection, that it is good for him to maintain this disposition if others do so as well. This is just what a successful argument for a Nash claim shows. We shall see that there are important differences between the Nash claim relied on in *TJ* and what I shall call *PL's Nash Claim*, but this fundamental point is not affected. The failure of Rawls's arguments for *TJ's Nash Claim* required Rawls to introduce a different Nash claim, and a different set of arguments for it.

The argument for *PL's Nash Claim*, like the congruence arguments in *TJ*, presupposes that members of the WOS have acquired a sense of justice. We have seen that in his earlier work, Rawls argued that the stability of the WOS is secured, in part, by the presence of various ideal-dependent desires. In §IX.2, we shall see that the Rawls of *PL* makes more explicit a point I have said he read into *TJ*: he describes the sense of justice itself as an ideal-dependent desire or as a set of ideal-dependent desires. In §§IX.3 and IX.4, I look at how the Rawls of *PL* thinks the sense of justice is acquired, an argument that is considerably more complicated than Rawls indicates. Because Rawls thinks the sense of justice is ideal-dependent, these two sections are concerned with the main topic of this chapter: the political ideals of justice as fairness. If justice as fairness is to be "as stable as one can hope for" (*TJ*, p. 399/350), then members of the WOS must judge, from within their comprehensive doctrines, that maintaining their ideal-dependent desires to be just belongs to their good. The argument for *PL's Nash Claim* shows that they would, at least under certain conditions. In §IX.5 I shall show that, contrary to what *TJ* leads us to expect, *PL's Nash Claim* does *not* support an argument for congruence.

Before I lay out *PL*'s argument for stability, let me say a word about the penultimate entry on my bulleted list. The account of stability I shall impute to Rawls in the bulk of this chapter is intended to show how a society well-ordered by *justice as fairness* could be stable. That is, of course, the case with which Rawls is primarily concerned. But beginning with the revised version of "Idea of an Overlapping Consensus," he conceded that members of a WOS might disagree about which conception of justice is most reasonable. A WOS was then described, not as a society in which everyone accepts a single conception of justice, but as one in which everyone accepts one or another member of a "class of liberal conceptions" (*PL*, p. 164). A just society characterized this way might seem to require still a different account of stability, since it might seem unlikely that people who disagree about justice would all want to live up to the ideals of justice as fairness. I believe this is correct, but that we will only be in a position to see why at the end of Chapter X.

§IX.1: *PL*'s Basic Argument for Stability

I have implied that I think it is possible to extract a main or central argument for stability from Rawls's later work. It will be useful to have that argument before us, and in this section I shall lay it out. Unfortunately, Rawls is not as clear as he might be about the structure of the argument. Some of his statements about stability in his later work can mislead about the conclusions he is trying to defend. Extracting the argument requires a certain amount of rational reconstruction; it may also entail some anachronism, since the argument brings together claims found in essays that were written some years apart. Even so, I do not believe that the extraction and reconstruction do violence to Rawls's texts. On the contrary, I think that the argument I shall sketch

makes the best sense of especially important passages in Rawls's later writings and of his transition to political liberalism, and that it is faithful to his most mature statements of his view.

One of the reasons that I think my reconstruction is faithful to Rawls's thought is that it brings to light similarities between his earlier and later treatments of stability. Recall that the argument for stability in *TJ* is a two-stage argument (see *TJ*, p. 453/397). The first stage consisted of showing that members of the WOS would normally acquire a sense of justice. The second consisted of showing that members of the WOS would judge that maintaining and acting from their sense of justice belongs to their good. In *PL* as in *TJ*, Rawls says that "stability involves two questions," the first of which is the question of whether members of the WOS would acquire a sense of justice (*PL*, p. 141). The argument I shall sketch in this section addresses the second of the two questions, which I take to be the question that is also addressed at the second stage of the stability argument in *TJ*—the question of whether members of the WOS judge that maintaining their sense of justice belongs to their good.

We saw that to clinch the argument for stability, the Rawls of *TJ* imagines an artificial perspective that members of the WOS can assume on their good. Someone who adopts this perspective gives "weight to [her] sense of justice only to the extent that it satisfies descriptions that connect it with reasons provided by the thin theory of the good" (*TJ*, p. 569/499). The Rawls of *TJ* identified a set of desires on which all members of the WOS would normally converge, desires the values of whose objects are given by the thin theory. He shows that these desires could best or only be satisfied by taking the desire to act from principles of justice as supremely regulative when others do. The common desires therefore give everyone in the WOS thin reasons to be just when others are. Rawls then offered "balance of reasons" arguments to show that those reasons are decisive. That is how he defended *TJ's Nash Claim*, C_N:

> C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

He then appealed to public knowledge of C_N to solve the *mutual assurance problem* and, eventually, to reach the *Congruence Conclusion.*

On my reading, Rawls follows a strategy in *PL* that is quite like the one he relied on in *TJ*. In *PL* as in *TJ*, Rawls imagines that members of the WOS all have a sense of justice, now acquired according to the process of social learning to be described in §§IX.3 and IX.4. I have already noted that in his later work, Rawls describes the sense of justice as a desire to act, not just from the principles of justice, but from the values and ideals of the political conception of justice. He then imagines members of the WOS asking themselves whether they and others have good reason to maintain and act from their sense of justice, just as he did in *TJ*. Rawls remarked in *TJ* that a WOS will be stable only if "the sense of justice that it cultivates and the aims that it encourages . . . normally

win out against propensities toward injustice" (*TJ*, p. 454/398). In *PL*, he says virtually the same thing, remarking that a WOS will be stable only if citizens' sense of justice "is strong enough to resist the normal tendencies to injustice" (*PL*, p. 142). Thus in Rawls's later work as in *TJ*, this question concerns each person's "balance of motives" (*TJ*, p. 454/398). In §VIII.5, I suggested that in his later work, Rawls shows somewhat greater concern with tendencies to or reasons for injustice that are rooted in citizens' identities and in their ethical views. We shall see later what some of those reasons are.

As in *TJ* so in *PL*, members of the WOS are assumed for purposes of this part of the stability argument not to give their sense of justice independent weight. And so in *PL*, he says that "citizens' overall views have two parts," one of which is the public conception of justice they endorse and the other of which is their comprehensive doctrine (*PL*, p. 38). In *PL*, Rawls seems to think that members of the WOS can, for purposes of argument, put aside their desire to live up to the public conception of justice as such or under that description, and adopt the viewpoint of their comprehensive doctrine—as he assumed in *TJ* that they can adopt the viewpoint of "a person following the thin theory" (*TJ*, p. 569/499). They can then ask whether their comprehensive views provide them reasons to maintain their sense of justice that are sufficiently weighty to counterbalance competing considerations.

Rawls argues, of course, that the answer is "yes," and so—given the way he describes a sense of justice in his later work—infers *PL's Nash Claim*:

> C_N*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

In the first part of *PL*'s basic stability argument, Rawls tries to establish this conclusion. How does he do so?

We have seen how the congruence arguments of *TJ* founder on the possibility that citizens who endorse some comprehensive views may reject the ideals of justice as fairness. The Rawls of *PL* therefore wants to show how citizens can follow their diverse, reasonable comprehensive doctrines while maintaining their sense of justice. And so on my reading, *PL*'s central argument for stability begins with the supposition that:

(9.1) Members of the WOS follow their comprehensive doctrines.

Of course, citizens do not always follow their comprehensive doctrines. Their religious or ethical views may make demands that they acknowledge, but fail to live up to. This kind of moral failure raises a set of problems that I shall not take up here. What matters for present purposes is that members of the WOS are assumed to want to live up to their comprehensive doctrines.

In *TJ*, the counterpart to (9.1) is the claim that members of the WOS would converge on certain desires, the value of whose objects is given by the thin theory of the good. Those are the desires referred to by C_3a, C_3b, C_3c, and C_3d—

the desires to express their nature, to avoid hypocrisy and deception, to live as friends, and to participate in social unions. Because members of the WOS converge on those desires, the Rawls of *TJ* was able to establish C_N by focusing on the typical member of the WOS, Joan. By assuming (9.1), Rawls does assume a convergence of desires, but that convergence is merely nominal. Members of the WOS all have the same desires *de dicto*, because they all want to live up to their comprehensive doctrines. But they do not all have the same desires *de re*, because their comprehensive doctrines differ. So the Rawls of *PL* cannot move from (9.1) to C_N* by asking what ends—understood *de re*—the "typical" member of the WOS desires. But if no member of the WOS can be treated typical, how can Rawls get from (9.1) to C_N* without asking about each person singly?

The crucial moves in Rawls's argument are his supposition that an overlapping consensus would obtain in a WOS, and his claims about what follows from that supposition. An overlapping consensus is a relation between "reasonable comprehensive doctrines likely to persist and gain adherents over time" in a WOS, taken together, and that society's public conception of justice (*PL*, p. 141). Rawls thinks that if an overlapping consensus obtains:

(9.2) "Reasonable doctrines endorse the political conception, each from its own point of view" (*PL*, p. 134).

(9.2) quotes what is probably Rawls's most familiar description of an overlapping consensus, and it fits a widely held picture of such a consensus—the picture I associated with the *Public Basis View* in §I.6. But talk of "endorse[ment]"—or of "fit" and "support" (*PL*, p. 145)—is somewhat vague. Elsewhere, Rawls is more precise. He says the possibility of an overlapping consensus is shown by the fact that

> the history of religion and philosophy shows that there 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 for a democratic regime. (*PL*, p. 140)

This suggests that when an overlapping consensus on a conception of justice obtains, then:

(9.3) Each comprehensive doctrine is "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 for a democratic regime."[1]

Appeal to an overlapping consensus at this point in the argument may be surprising, since it is sometimes thought that the Rawls of *PL* just took the

1. *PL*, p. 169; also Rawls, "The Domain of the Political and Overlapping Consensus," *Collected Papers*, pp. 473–96, p. 485.

problem of showing stability to be the problem of showing that there would be or could be an overlapping consensus on justice as fairness. As we shall see, there are passages that seem to support this reading, particularly in Rawls's earliest discussions of an overlapping consensus. Since at this point, I am simply trying to lay out what I take to be the basic stability argument, I shall put off discussion of this alternative reading until §IX.5.

I am confining my inquiry to an overlapping consensus on justice as fairness rather than on a family of liberal political conceptions. The "values appropriate to the special domain of the political as specified by the political conception of justice for a democratic regime" are therefore the political ideals of justice as fairness and the values of realizing them. These are the political values and ideals referred to by C_3'. So (9.3) implies that:

(9.4) Each comprehensive doctrine is "either congruent with, or supportive of, or else not in conflict with" the political ideals referred to by C_3' and the values of realizing them.

The three cases "congruent with," "supportive of," and "not in conflict with" differ in ways that will prove significant later, but for purposes of laying out the main lines of Rawls's argument those differences can be put aside. For now, suffice it to say that (9.4) refers to the convergence on values and ideals on which I said Rawls relies for stability. For when any one of these relations holds, then:

(9.5) According to each comprehensive doctrine, the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them" (*PL*, p. 156).

We saw that in *TJ*, when Rawls considered the weightiness of thin reasons to be just, he was explicit about what considerations needed to be outweighed. They were considerations that tell in favor of acting unjustly, such as the desire for money that can be saved by cheatings on taxes. The Rawls of *PL* has to show that considerations drawn from comprehensive doctrines will outweigh those considerations as well. But his appreciation for the pluralism of comprehensive doctrines broadened his concern to encompass an additional set of considerations that may conflict with the values of justice as fairness. By *PL*, they include the value of what can be gained by acting unreasonably.

This broadening of concern is important, since unreasonable action includes considerably more than the kinds of shirking and free-riding that I used to exemplify unjust action in my discussion of *TJ*. Unfortunately, Rawls is not as explicit about this as he might be. In one later work, he does seem to equate the question of how the values of the political can normally outweigh conflicting values with the question of how someone can maintain a comprehensive doctrine and yet not think he can use political power to enforce it.[2] It would be a mistake to equate

2. Rawls, *Restatement*, p. 189.

the two questions. But the attempt to enforce aspects of one's comprehensive doctrine—by, for example, arguing for political measures simply on the grounds that the policies to be enforced are supported by that doctrine—is an important example of an action that is contrary to the political ideals of justice as fairness, despite the fact that it is not a straight-forward example of free-riding or injustice. Rawls's use of the example testifies to the broadening of concern to which I referred in the last paragraph. I shall assume that the value of enforcing aspects of one's comprehensive doctrine politically – like the value of what can be gained by garden varieties of injustice—exemplifies the values that the Rawls of *PL* thinks are contrary to justice as fairness and need to be outweighed.

(9.1) says that citizens follow their comprehensive doctrines. Citizens who follow a comprehensive doctrine follow it by ordering values as their comprehensive doctrine says they are to be ordered. That, we might think, is what it is to follow a comprehensive doctrine. So (9.5), together with (9.1), would seem to have *some* implications for the judgments of political value reached by members of the WOS. But what implications?

If a reasonable comprehensive doctrine "endorse[s]" justice as fairness, as (9.2) says such doctrines would in a WOS, that does not mean that it endorses justice as fairness unconditionally. Rather, it endorses it as the conception of justice that is suited, perhaps best suited, to play a shared, public role—the role of adjudicating competing claims among citizens who follow it willingly. Citizens who follow their comprehensive doctrine endorse justice as fairness, or give its values certain weight, subject to the same condition. They endorse justice as fairness as the conception they should act from when everyone else does too. So what (9.1) and (9.5) imply is that:

(9.6) Each member of the WOS judges, from within her comprehensive view, that the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them," at least when others reach the same judgment.

But what does Rawls mean by "outweighing"?

Comprehensive doctrines are sources of reasons for those who adhere to them, and when Rawls says that some values outweigh others, I take it that what he means is that some are taken to be sources of more compelling or weightier reasons for action. So I take it that from (9.6), Rawls infers *PL's Nash Claim*:

C_N*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

This conclusion can be illustrated using the following payoff table, where A > B, C, and D.

Table IX.1

	Player 2	
	Maintain desire to live up to values and ideals of justice as fairness	Decide case-by-case
Maintain desire to live up to values and ideals of justice as fairness	A, A	C, B
Player 1		
Decide case-by-case	B, C	D, D

When I illustrated *TJ's Nash Claim* C_N using similar payoff tables, I said that the payoffs for maintaining a regulative desire to act from principles of justice had to be reckoned using the thin theory. *PL's Nash Claim* C_N* does not refer to judgments made from within the thin theory. But as I have already suggested, it does refer to a viewpoint in which the desire to be just as such, and cognate desires, are left out of account. And so in Table IX.1, the payoffs for living up to the values and ideals of justice as fairness—A and C—are reckoned on the basis of comprehensive doctrines, without taking account of the value of being just.

I argued that the Rawls of *TJ* thought that justice as fairness, when institutionalized, would bring about the truth of *TJ's Nash Claim*; that was an important part of the case for inherent stability. We shall see in Chapter X that the Rawls of *PL* thinks justice as fairness, when institutionalized, would bring about the truth of *PL's Nash Claim*. My argument for this conclusion gets some support from Rawls's very telling remark in *PL* that "a reasonable and effective political conception may bend comprehensive doctrines toward itself" (*PL*, p. 246).

We saw that *TJ's Nash Claim* incorporates what I called a "reciprocity rider." It says members of the WOS would maintain their sense of justice as supremely regulative in a special case, the case in which the plans of others are similarly regulated. Because C_N includes this rider, Rawls needed to confront the *mutual assurance problem*. His solution to that problem, together with C_N, enabled him to infer:

> C_6: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

And we saw how he moved from C_6 to a conclusion about how members of the WOS would judge their reasons from the viewpoint of full deliberative rationality, the *Congruence Conclusion*:

C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her desire to act from the principles of justice as a highest-order regulative desire in her rational plans.

The Rawls of *PL* wants to proceed similarly. As Table IX.1 shows, *PL's Nash Claim* C_N^* says that each person takes a life in which he acts from the values and ideals of justice as fairness to be his best reply to the similar plans of others, when judged from within comprehensive doctrine. So *PL's Nash Claim*, like *TJ's*, includes a "reciprocity rider" and the *mutual assurance problem* still needs to be solved. How does Rawls solve it?

The WOS is and is known to be a just society. So to solve the *mutual assurance problem*, what each person in the WOS needs to know is that, as Rawls put it in "Domain of the Political and Overlapping Consensus," "no reasonable and rational person in the well-ordered society of justice as fairness is moved by rational considerations of the good not to honor what justice requires."[3] Now suppose not just that an overlapping consensus obtains in a WOS, as I assumed between (9.2) and (9.3), but that everyone in the WOS knows that it obtains. Then each citizen is in a position to know that

(9.5) According to each comprehensive doctrine, the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them" (*PL*, p. 156).

In that case, each person knows that, at least under normal circumstances, no one else's comprehensive view provides sufficient "considerations of the good not to honor what justice requires" and, indeed, each knows that other comprehensive views normally provide reasons *to* honor what justice requires. More intuitively put, each person knows that not only does no one's view of what is good in life provides him sufficient incentive to cease being just, but each knows that others' views normally provide them incentives to continue to be just. So everyone knows that the condition imposed by the reciprocity rider is satisfied. In that case, Rawls could infer a claim he explicitly makes in "Reply to Habermas," a claim I shall call "C_9":

C_9: "citizens will judge (by their comprehensive view) that political values either outweigh or are normally (though not always) ordered prior to whatever nonpolitical values may conflict with them" (*PL*, p. 392)[4].

It tells in favor of my reading that Rawls not only defends this conclusion, but says that it depends upon just the suppositions I have made. For it depends, he

3. Rawls, "Domain of the Political," *Collected Papers*, p. 487, note 30.

4. Rawls actually prefaces the quoted remark with the phrase "we hope". The reasons for the phrase are given at *PL*, p. 392 note 29 and I shall ignore them here.

says, on "the existence and public knowledge of a reasonable overlapping consensus" (*PL*, p. 392). I still need to show why Rawls is justified in thinking that an overlapping consensus would obtain and would be known to obtain in a WOS. I shall do so in Chapter X.

Of course, as in *TJ* so in *PL*, what Rawls really wants is a conclusion about how each person would judge his balance of reasons in the viewpoint in which he actually draws up his plans of life. And so he wants to move from C_9 to:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

The truth of C_9 shows that when members of the WOS ask themselves what their comprehensive doctrines demand, they do not normally see any conflict between those demands and the demands of justice. Indeed, it shows that their comprehensive doctrines normally pull them toward being just persons. The only difference between the viewpoint of comprehensive doctrine and the viewpoint of full deliberative rationality is that in the former, we leave out of account our desires to do justice for its own sake. If our balance of reasons tips toward justice—if we are pulled toward justice—even when those desires are left out of account, the balance will not shift when they are taken into account. So just as the Rawls of *TJ* could move from *TJ's Nash Claim* via C_6 to C_C, the Rawls of *PL* can move from *PL's Nash Claim* via C_9 to C_{PL}. Like the *Congruence Conclusion* C_C, C_{PL} describes an equilibrium state, a state which would be stabilized by the enduring character of the forces that bring it about.

We have seen that in *TJ*, Rawls wanted to show that members of the WOS would develop and maintain a desire to regulate *all* their other desires by their desire to act from the principles of justice. For reasons I shall discuss in §IX.5, the Rawls of *PL* does not want or need so strong a conclusion. For him, it is enough to show that each person would develop and maintain a desire to treat the principles and values of justice as fairness as regulative of political life. We shall see that since citizens of the WOS can realize the goods and ideals of justice as fairness only by giving the political conception that kind of authority, it is enough to show that they would acquire a sense of justice and that—as we shall see—C_{PL} is true.

My reading of Rawls's later treatment of stability gains some credence from the fact that, if it is right, there are a number of similarities between his earlier and later treatments—similarities that have, I think, gone largely unnoticed. I have already drawn attention to some of those similarities. Let me now mention two more.

First, we saw that in the original *Deweys*, members of the WOS were shown to have what the Rawls of *PL* called "conception-dependent desires" to live up to the ethical ideals of conduct, friendship, and association included in justice as fairness. These are the desires referred to by C_3 and C_3*. At the end of §VIII.1 and in §VIII.2, we saw why Rawls thought members of the WOS would

have those desires, whatever their comprehensive doctrine. We have also seen that they can satisfy those desires only by living justly, and the presence of these desires account for why members of the WOS act justly in ordinary life. The Rawls of *PL*, like the Rawls of *TJ* and the original *Deweys*, thinks that members of the WOS have conception-dependent desires to live up to the ideals of justice as fairness. In *PL*, those are not the desires to live up to *ethical* ideals, referred to by C_3 or C_3*. They are the desires to live up to *political* ideals, referred to by C_3'. In §III.3, we saw where Rawls's earlier treatment of stability appealed to C_3 or C_3*. In §IX.2, I shall show where the later treatment appeals to C_3'. For now, note that like the ideals referred to by C_3 and C_3*, the ideals referred to by C_3' can be realized only if members of the WOS are just persons. And so the Rawls of *PL*, like the earlier Rawls, thinks stability depends upon the presence of ideal-dependent desires.

Second, if members of the WOS have the ideal-dependent desires that C_3 asserts, as the early Rawls argued that they would, then their plans of life must include the satisfaction of those desires. Since they can satisfy those desires only if they are just, congruence follows immediately if members of the WOS take those desires into account when they ask whether being just belongs to their good. Indeed, the immediacy of the implication is such that Rawls thought an argument from C_3 or C_3* to congruence would be trivial. Moreover, Rawls wanted to show that in the WOS, no one's desire to be just, or to live up to the ideals referred to by C_3 and C_3*, would be undermined by her other desires, and that each member of the WOS would have that assurance about all the others. To show this, he tried to argue for C_6.

The ideals to which C_3' refers are political rather than ethical ideals. They guide "public life" (*PL*, p. 77) rather than life as a whole. Members of the WOS cannot be sure that others' desire to live up to a political conception of themselves or their society will be effective, or will help to stabilize a WOS, without knowing that those desires will not be undermined by desires that make competing demands. The weight someone attaches to these competing demands might lead her to repudiate or to compromise her sense of justice. Until we see how members of the WOS relate political ideals and values to other ideals and values, and what they have reason to think about the conduct of others, it will not be at all clear that members of the WOS judge that their good includes satisfying the desires referred to by C_3'. The problem of showing that members of the WOS would maintain their sense of justice therefore depends upon showing C_9. It depends upon showing that members of the WOS, judging according to their comprehensive views, would think they have sufficiently weighty reason to be just over time. Thus the Rawls of *PL*, like the Rawls of *TJ*, uses a "balance of reasons" argument to show that members of the WOS would judge that maintaining their sense of justice belongs to their good.

My reading of Rawls's political turn therefore shows an underlying similarity of strategy and concern between his two treatments of stability. The similarities I have highlighted give my reading some credence. As we shall see, it gains further credence from the fact that it accounts for the differences

between *TJ* and *PL* that I have said need to be explained. But my reading also faces some textual obstacles that I shall have to confront.

One of those difficulties concerns the way I have said the Rawls of *PL* sets up the problem of stability. I said above that in *PL* as in *TJ*, that problem is explicitly said to involve two questions, the first of which is whether members of the WOS would acquire a sense of justice (*PL*, p. 141). But in *PL*, the second question is never said to be what I have taken it to be: the question of whether members of the WOS would judge that maintaining their sense of justice belongs to their good. Rather, in the place in *PL* where the two questions of stability are most clearly distinguished, the second question is said to be—rather than to depend upon—the question of whether the political conception can be the focus of an overlapping consensus (*PL*, p. 141). As if to underline the differences between his earlier and later treatments of stability, Rawls virtually drops the word "congruence" and its cognates from his lexicon—the exception being its occurrence in step (9.3). Moreover, the Rawls of *PL* never says that he relies on the strategy I ascribe to him, nor that he wants to show stability by a "balance of reasons" argument. *PL* does not contain even a passing reference to "the hazards of the generalized prisoner's dilemma" (*TJ*, p. 577/505). By the time Rawls wrote "Reply to Habermas," he seems to answer the two questions about stability indirectly, by arguing that justice as fairness can enjoy various kinds of justification (*PL*, pp. 385ff).

I believe that these exegetical obstacles can be overcome, and my interpretation can be sustained. But sustaining it is not a matter of interpreting a few continuous pages of text. Rather, it requires pulling together a number of crucial passages from very different places in Rawls's later works. I will try to defend my claims about *PL*'s treatment of stability, beginning with my claims about the way the Rawls of *PL* sets up the problem. I shall then be in a position to show why Rawls thought an overlapping consensus would obtain in a WOS and to return in Chapter X to the argument I have sketched in this section.

§IX.2: C_3' and the Sense of Justice

In my introductory remarks, I said that Rawls's later account of stability relies on:

> C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and society included in justice as fairness.

I have not, however, said where in the account Rawls appeals to this claim or argues for it.

At the beginning of the previous section, I noted that in *PL* as in *TJ*, the treatment of stability has two parts. The first part is an argument that members of the WOS would normally acquire a sense of justice. In §§III.2 and VIII.2, I claimed that Rawls's treatment of the sense of justice underwent a significant development after the publication of *TJ*, so that ideal-dependent

desires became central to it. If my claim is right, then the first part of *PL*'s discussion of stability is an argument for C_3'. The second part of that discussion, like the second part of *TJ*'s discussion, then shows that members of the WOS would all plan to maintain their sense of justice. To substantiate my claim about the sense of justice, I now want to look more closely than I have so far at how Rawls's treatment of that sentiment developed between *TJ* and *PL*.

In *TJ*, Rawls had said that the sense of justice is "a normally effective desire to apply and to act upon the principles of justice" (*TJ*, p. 505/442). This definition and *TJ*'s identification of the last stage of moral development with the morality of principles convey the impression that, in *TJ*, the sense of justice is what the Rawls of *PL* would call a "principle-dependent" desire.[5] Even as late as the original *Deweys*, Rawls still defines a sense of justice as "the capacity to understand, to apply and to act from . . . *the principles of justice*."[6] But beginning at least in "Political not Metaphysical,"[7] and through *PL*, the capacity for a sense of justice is consistently said to be, not a capacity to apply and act on *the principles*, but "a capacity to understand, to apply and to act from *the public conception of justice* which characterizes the fair terms of social cooperation" (*PL*, p. 19, emphasis added).[8] The latter description of a sense of justice is implied in Rawls's earlier work, including work that immediately postdates *TJ*,[9] but there the difference between acting on principles and acting on a conception is elided.[10]

In fact, the differences are significant. A public conception of justice includes the principles of justice, of course. But it also includes considerably more. In the case of justice as fairness, it includes significant theoretical apparatus, such as the OP, as well as political ideals and values. So according to the more expansive definition given in *PL*, someone who has a developed sense of justice informed by justice as fairness wants to act from the principles. She also wants to live up to the ideals justice as fairness includes and to realize its values. Most important for present purposes, she wants to live up to its ideals of conduct and to be the sort of person who consistently gives the principles of right the appropriate place in political reasoning. In §VIII.2, I argued that in *TJ* and certainly by the original *Deweys*, ideal-dependent desires are an important

5. So too does the parallel Rawls draws between the sense of justice and the sense of grammaticalness (*TJ*, p. 41).

6. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 312.

7. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 398.

8. *PL*, p. 302 describes the sense of justice as a desire to act from principles, but the description occurs in "Basic Liberties and Their Priority." That essay was first published in 1982 and was reprinted unchanged in *PL*. It therefore antedates "Political Not Metaphysical."

9. See, for example, Rawls, "Reply to Alexander and Musgrave," *Collected Papers*, p. 233.

10. And so at "Reply to Alexander and Musgrave," *Collected Papers*, p. 233 Rawls says a WOS is a society in which "everyone accepts, and knows that others accept, the same principles (the same conception) of justice."

part of a sense of justice. In Rawls's later work, such desires are not just part of a sense of justice, they are central to it. This is confirmed by an important section in the revised *Deweys*. There Rawls distinguishes principle- and conception-dependent desires and lays out the "reasonable moral psychology" by which the sense of justice is acquired. He says quite clearly that "for us these [conception-dependent desires] are the most important" (*PL*, pp. 83–84).

Why did Rawls develop his view in this way?

One reason for the development, I believe, is that Rawls came more fully to appreciate a line of thought he found in Kant.[11] The just person is moved by principles of right to do justice for its own sake. But principles of right, at least those that apply to the basic structure, are highly abstract and may not move us very powerfully by themselves. Our desire to act from them is more effectively elicited by seeing how they can be exemplified in a just society and in the lives of such a society's members. I have stressed that ideals or conceptions are relatively specific: they specify concepts, such as the concept of a free and equal person. The ideals of justice as fairness—and the connections among them—are specific enough that, when members of the WOS are made aware of them, they see what they and their society can be if they act from the principles of justice. Since this elicits their desire to act from the principles of justice, it is natural to describe that desire as ideal- rather than principle-dependent. The premises of this line of thought—the claims about how our moral motivations are most effectively elicited—are not entirely absent from *TJ*, but they are hardly central.[12] I believe Rawls eventually recognized the need to develop this line of thought further, and followed it to the natural conclusion that the sense of justice is ideal-dependent.

Another reason for the development can be brought to light by thinking about what it means to say, as I have, that ideal-dependent desires are "central" to a sense of justice. I do not mean that the desire to be a certain kind of person is the characteristic motive of the just person, if the characteristic motive of some kind of action is understood to be the motive from which those actions are typically or characteristically done. Nor do I mean that it is the motive from which an *ideally* just person would typically act. The claim that the ideal-dependent desire is central is compatible, as it must be, with the claim that just persons—even ideally just persons—typically are moved by principles and perform just acts for their own sake. Of course, the just person's desire for full autonomy, for example, may play an important role in her practical deliberations, but this role does not explain the centrality of the ideal

11. Here I refer to Rawls, *Lectures on the History of Moral Philosophy*, pp. 201–2 and 212–14.

12. I am thinking of the important but unelaborated remark about "the ideal of persons" at *TJ*, p. 478/419 as well as Rawls's remark at *TJ*, p. 477/418: "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."

either. Rather the appropriateness of giving the ideal such a role is part of what needs to be explained.

For Rawls, the sense of justice is a family of beliefs, desires, and dispositions. As we have seen, he thinks the object of the desire to be just is liable to a *diversity of descriptions* and that the sense of justice is connected with natural attitudes and dispositions (*TJ*, pp. 485ff/425ff). Thus if the sense of justice is a family, it is a large and extended one. Rawls's reliance on the *diversity of descriptions* and on the connections between moral and natural attitudes to establish congruence shows how he uses the size and extension of the family to advantage. But though Rawls treats them as advantages, these features of the family raise the question of just what relates its members, so that the sense of justice constitutes a single moral sentiment. What kind of unity would provide an acceptable answer to that question?

I believe Rawls wants to show that the sense of justice exhibits what we might call a "rational unity." More precisely, I believe he thinks the constituents of a sense of justice are united into a single coherent sentiment by a unified rationale which explains why we want the object of a sense of justice under the various descriptions of which it admits, and why we value the various affective dispositions Rawls connects with the sense of justice. But what can that rationale be?

I mentioned a moment ago that even in *TJ*, Rawls seems to think that a sense of justice includes the desire to live up to ideals of the person and society. But the Rawls of *TJ* also seems to think that when fully developed, the sense of justice is unified by principles of right, for he says that the morality of principles is "the last stage at which all the subordinate ideals are finally understood and organized into a coherent system by suitably general principles" (*TJ*, p. 478/419).

The problem with this claim is that, as we saw, the principles of justice satisfy a *diversity of descriptions*: for example, they are described as the commonly recognized morality of a WOS, as the principles regulating a social union of social unions and as principles which are such that, by acting from them for their own sake, we express our nature. The problem with *TJ*'s claim about how the principles "organize" "subordinate ideals" into "a coherent system" is that—though it may indeed "organize" "subordinate ideals"—it still seems to leave us without a coherent and systematic rationale for acting from the principles under these various descriptions.

The absence of any such rationale would be something of an embarrassment for Rawls. For as we saw in §V.1, Rawls thinks that the availability of a *diversity of descriptions* gives contract theory an advantage over intuitionism, since the latter has to treat the desire to be just as a preference on a par with the preference for tea over coffee precisely because it cannot offer alternative descriptions of that desire which connect it with "our rational aims" (*TJ*, pp. 476–77/417–18). The need for some unified rationale that shows the point of acting on principles of justice became more pressing in *PL*, where Rawls gave a prominent role to other principles that would also be chosen in the OP and that govern the ways in which the principles of justice are applied—namely, the

guidelines of public reason and the liberal principle of legitimacy. Is there a single rationale, or "rational aim," that shows the point of regulating all one's desires by the desire to act from the many principles that would be adopted in the OP, and from the principles of justice under their *diversity of descriptions*?

I believe Rawls came to think that a rationale for regulating political reasoning by all these principles is provided by the ideals of a free and equal person and of society as a fair scheme of cooperation, as those ideals are specified by justice as fairness. For only someone who has a settled disposition to treat the principles, under their various descriptions, as regulative of her political reasoning is a good and reasonable citizen of a just society. Principles of right continue to enjoy one kind of priority over those ideals: as I'll say in more detail at the beginning of the next section, the ideals of justice as fairness are specified from more basic or rudimentary conceptions of the person and of cooperation by appeal to the principles of right. This is what we might call the "conceptual priority of the principles." But since the just person of *PL* wants to realize those ideals, it is the ideals, once specified, that display the "rational aims" (*TJ*, p. 476/417) to which the principles are related. And so, while principles may enjoy "conceptual priority," ideals enjoy what we might call "rational priority." A moment ago, I said that the Rawls of *TJ* thought the morality of principles is "the last stage at which all the subordinate ideals are finally understood and organized into a coherent system by suitably general principles" (*TJ*, p. 478/419). I believe the Rawls of *PL* would reverse this description of the last stage of moral development, and say that it is the stage "at which [principles] are finally understood and organized into a coherent system [by the ideals of justice as fairness]."

It should now be clear where Rawls's later account of stability appeals to:

C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and society included in justice as fairness.

For the sense of justice, as described in *PL*, is or includes the desires to which C_3' refers. The argument for C_3' thus answers the first of two questions that stability is said to involve: the question of whether members of the WOS would acquire that moral sentiment. Thus what needs to be shown by the account of moral development at work in *PL* is not that members of the WOS would acquire the morality of principles. What needs to be shown is how they acquire the desire to live up to the ideals included in justice as fairness—in *PL*, the ideals to which C_3' refers. Rawls does in fact show this, in a set of arguments I shall reconstruct in the next two sections. Because the argument I sketched in §IX.1 answers the second question, it presupposes that the first question has been answered affirmatively and that C_3' is true.

§IX.3: C_3' and the Ideals of Conduct

I now want to look at how the Rawls of *PL* defends C_3', which the arguments of the last section show to be the claim that members of the WOS would

normally acquire a sense of justice. Because C_3' refers to political rather than ethical ideals, it is weaker than the claims on which his earlier treatment of stability depended:

C_3: All members of a WOS want to live up to the ideals of personal conduct, friendship, and association included in justice as fairness.

and

C_3^*: All members of a WOS want to live up to the ideal of full autonomy.

The relative weakness of these claims is deliberate for, as we have seen, the strength of C_3 and C_3^* is part of what undid Rawls's earlier account of stability. Weakening C_3 and C_3^*, and defending the weaker C_3', required significant changes in justice as fairness, and looking at how Rawls defended C_3' explains some of the changes between *TJ* and *PL*. My examination of the defense of C_3' is divided into two parts. In this section, I will look at how Rawls defends the claim that members of the WOS want to live up to a political ideal of conduct. In §IX.4, I will look at how he defends the claim that they want to live up to political ideals of interpersonal life.

To see just how and why Rawls weakened C_3 and C_3^*, it helps to recall how he moved from the concept of the person to the ideal of conduct to which C_3 and C_3^* refer. That move was possible because the tradition of democratic thought inherited the concept of the person "understood as the concept of someone who can take part in, or who can play a role in, social life and hence exercise various rights and duties".[13] The tradition specified that concept somewhat, into a conception of a free and equal rational person. *TJ* took over this more specific conception from the democratic tradition and asked, in effect, how that conception could best be lived out or expressed in the lives of citizens. It used the principles of justice to answer that question, specifying or "articulat[ing]" (*PL*, p. 84) the conception still further and arriving at the ideal referred to by C_3 and C_3^*.

It is because Rawls specified a conception of the person that C_3 and C_3^* refer to an ideal of *personal* conduct. The ideal is realized in the whole lives of persons; it is an ideal they live up to when they regulate their plans by the principles of justice. The ideal of conduct referred to by C_3' is not an ideal of personal conduct so understood. Rather, it is an ideal of democratic citizenship. It is arrived at by beginning with a concept of the *citizen* as "someone who can take part in, or who can play a role in, social life and hence exercise various rights and duties." The democratic tradition specifies *that* concept into a conception of the free and equal rational citizen. In *PL*, Rawls asks, in effect, what the best way is to express or live out *this* democratic conception. He uses the principles of right further to specify the conception and arrive at "the ideal of

13. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 397.

citizenship as characterized in justice as fairness" (*PL*, p. 84). *That* is the ideal of conduct *PL* includes in justice as fairness and to which C_3' refers.

In Rawls's later work, he says explicitly that justice as fairness begins from "basic intuitive ideas" that are "implicit in the public culture of a democratic society."[14] This is sometimes thought to mark a difference between *TJ* and *PL*. I argued in §III.1 that while Rawls's explicitness about this marks a difference between his earlier and his later work, Rawls always began within the liberal democratic world. In *TJ* as in *PL*, he began from conceptions he found in the democratic tradition.

A real difference between *TJ* and *PL* is the scope of conduct that the resultant ideals are supposed to guide. The ideal of conduct in *PL* is to guide members of the WOS insofar as they act in the role of citizen, and it is an ideal they can live up to in that part of their conduct. Thus, an important element of "the ideal of citizenship" is full autonomy. Whereas in the original *Deweys* Rawls wrote that members of the WOS realize full autonomy "in their daily lives,"[15] in the revised version included in *PL*, he says that they realize full autonomy "in public life." (*PL*, p. 77) Another way members of the WOS live up to the ideal of full autonomy is by showing "respect for the precepts governing reasonable political discussion."[16] These precepts are, or include, the guidelines of public reason. Like the principles of justice, they guide members of the WOS in specifically public life, for they apply to certain debates in the public political forum. They do not govern discussion—even political discussion—elsewhere.[17]

Rawls therefore specifies the ideal of conduct—and, as we shall see, other ideals as well—using two different sets of principles. Each set of principles is used to specify an ideal that is constitutive of the more inclusive ideal of citizenship. The principles of justice help to define the ideal of full autonomy; the precepts governing reasonable political discussion help to define what Rawls calls "the ideal of public reason."[18] It is clear how the ideal-dependent desire to be fully autonomous bears on stability, since members of the WOS can satisfy that desire only if they are just. The bearing of the ideal of public reason on stability may be less clear, since it is not clear that someone can realize that ideal only if she is just. We shall see the connection shortly.

Rawls says less than he might about just where the boundaries of political discussion and public life lie. I am less concerned with locating them than I am with the ideas underlying Rawls's insistence that there *are* such boundaries, boundaries that carve out what he calls "the domain of the political." One of those underlying ideas is that when members of the WOS live up to the

14. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 396.
15. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 315.
16. Rawls, "Domain of the Political," *Collected Papers*, pp. 484–85; see also *PL*, p. 139.
17. Rawls, "The Idea of Public Reason Revisited," *Collected Papers*, pp. 573–615, p. 576.
18. Rawls, "Public Reason Revisited," *Collected Papers*, pp. 576–77.

ideal of citizenship, they realize goods—such as mutual respect among citizens and full autonomy—the values of which do not depend upon nonpolitical values. Another is that the ideal of citizenship need not guide conduct in the whole of life. Insofar as members of the WOS act in other roles, they can be guided by other ideals of conduct. This means that to realize the ideal of conduct that *PL* includes in justice as fairness—understood as including full autonomy—they need not take the desire to act from principles they give themselves as regulative of the entirety of their plans of life. They need only take the desire to act from principles of right as regulative of their political lives, however these are finally delineated.

In §VIII.2, we saw that according to the original *Deweys*, the desire to live up to the ideal of full autonomy referred to by C_3 and C_3* is elicited when the justification of the principles of justice and their application are public knowledge. The public justification of the principles includes something like what I have called the Pivotal Argument. We have seen that that argument relies on:

(1.1) We are free and equal rational persons who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

It also relies on:

(1.9) The OP is a choice situation in which our nature is the decisive determining element.

So if the Pivotal Argument is public knowledge and is publicly accepted, then members of the WOS accept (1.1) and (1.9). If they accept what I called the *KI Claim*—the claim that "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*, p. 253/222)—then they see that acting from principles they would give themselves in the OP is the best expression of their nature as free, equal, and rational. Seeing this elicits the desire to be the kind of person who acts from those principles, a person who realizes the ideal of personal conduct in justice as fairness.

C_3' does not imply that members of the WOS want to live up to an "ideal of the *person*." It implies, rather, that they want to live up to an ideal of *citizenship*. The Rawls of *PL* does in fact argue that they would have this ideal-dependent desire. That he argues for this conclusion goes some way to showing that he argues for C_3'. *PL*'s argument that members of the WOS want to live up to an ideal of citizenship parallels the earlier argument that they want to live up to an ideal of the person, particularly in the crucial educative role it gives to the publicity condition and the original position. Looking at the details of the argument sheds light on another of the significant changes between *TJ* and *PL*: the claim that OP represents the powers of citizens rather than persons.

In the version of *Deweys* included in *PL*, Rawls argues that the desire to live up to an ideal of citizenship rather than to an "ideal of the person" is what

is elicited when the justification of the principles and of their applications are fully public. The revised version of the passage from the original *Deweys* that I quoted in §VIII.2 says that when the full publicity condition is satisfied:

> a political conception assumes a wide role as part of public culture. Not only are its first principles embodied in political and social institutions and public traditions of their interpretation, but the derivation of citizens' rights, liberties and opportunities *invokes a certain conception of citizens as free and equal*. In this way, citizens are made aware of and educated in this conception [or ideal]. They are presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain. To realize the full publicity condition is to realize a social world within which the *ideal of citizenship* can be learned and may elicit an effective desire to be that kind of person. (*PL*, p. 71, emphases added)

The logic of this argument is the same as that of the corresponding argument from the original *Deweys*: members of the WOS all know that they and others benefit from their basic institutions, relative to a suitable benchmark of comparison. Their benefits are justified by public appeal to a certain conception of themselves. Public justification thus presents them with "a way of regarding themselves that otherwise they would most likely never have been able to entertain." Knowing that they and others benefit from being regarded that way, and finding this view of themselves attractive, they accept that view of themselves and desire to live up to it. When they fulfill that desire, they express their free and equal citizenship.

One way in which this argument departs from the parallel argument in the original *Deweys* is that Rawls now explicitly states a psychological premise on which the last step depends. He says that members of the WOS "want to be, and to be recognized as, [fully cooperating] members" of their society (*PL*, p. 81). This desire is part of their "moral sensibility" (*PL*, p. 81). The public conception of justice contributes to their moral education by specifying and presenting them with the object of this desire.

To sustain this claim, Rawls would need to show that when the justification of the principles is fully public, members of the WOS are educated in an "ideal of citizenship." It is not clear that he could show this if the public justification of the principles appealed to (1.1) and (1.9). For these are not claims about an ideal of citizenship. They are claims about the nature of persons and its representation in the OP. The conception of the person they appeal to was to be used in the formation and justification of government policies.

Suppose instead that the public justification of the principles and their application appealed to what I shall call the *political analogues* of (1.1) and (1.9):

(1.1') We are free and equal rational citizens who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

(1.9') The OP is a choice situation in which our nature as citizens is the decisive determining element.[19]

Suppose further that everyone in the WOS accepted the *political analogue* of the *KI Claim*:

> to express one's nature as a citizen is to act on the principles that would be chosen if this nature were the decisive determining element.

The two principles of justice would then have a new public basis. That basis would be a revised version of the Pivotal Argument, which results from replacing (1.1) and (1.9) with (1.1') and (1.9') and replacing the word "persons" with "citizens" in the other steps where it occurs. Members of the WOS could then see that acting from and applying principles they would give themselves in the OP is the best way to express their citizenship. Seeing this, Rawls thinks, would elicit the ideal-dependent desire to act from the principles in public life.

The restriction of the ideal to one of citizenship, lived out in public life, is supposed to make it possible for members of the WOS to acquire the desire even if they want to live up to other ideals, associated with their comprehensive views, in other areas of life. Thus by relying on the *political analogues*, the Rawls of *PL* thought he could do what he had tried but failed to do in *TJ*: found the principles of justice on a publicly endorsed view of themselves that everyone in the WOS could aspire to live up to.

Two of the most salient differences between the Rawls of *TJ* and the later Rawls are that the later Rawls begins with a specifically "political conception of the person"[20] and insists that the OP represents the nature of citizens rather than of persons. I have tried to pin down the claims in which those differences find expression and to show what a difference in Rawls's arguments the reliance on these premises is supposed to make. The claims I have identified are the *political analogues* of (1.1), (1.9), and the *KI Claim*, which—unlike (1.1), (1.9), and the *KI Claim* themselves—are claims about the nature of citizenship and its representation in the OP. While reliance on these claims reshaped the Pivotal Argument, it is important to see why Rawls thought that the argument had to be reshaped. The problem with the original version was not that members of the WOS might disagree with its premises, as the *Public Basis View* holds. Rather, the problem was that the original version of the Pivotal Argument could not play the necessary role in bringing about inherent stability. It could not, when publicized, elicit the ideal-dependent desires needed to stabilize justice as fairness. Two very important changes between *TJ* and *PL*

19. It may be objected that artifacts such as social roles cannot have natures, and that talk of "our nature as citizens" is therefore too philosophically confused to impute to Rawls. But note that at *PL*, p. 203 he says that "having the two moral powers" is "part of the essential nature of citizens."

20. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 397, note 15.

can therefore be explained as responses to shortcomings Rawls found in his earlier treatment of stability.

Readers often wonder how much of *TJ*'s account of moral development survives the transition to *PL*.[21] The story I have just drawn from the revised *Deweys* would seem to force changes in the account of moral development in *TJ*, which culminates in the acquisition of a morality of principles rather in the acquisition of a set of ideal-dependent desires. Yet even in *TJ*, when Rawls asks about how someone develops a morality of principles, he slides almost immediately to how ideal-dependent desires are acquired. He says:

> I should now like to consider the process whereby a person becomes attached to these highest-order principles themselves, so that just as during the earlier phase of the morality of association he may want to be a good sport, say, he now wishes to be a just person. The conception of acting justly, and of advancing just institutions, comes to have for him an attraction analogous to that possessed before by subordinate ideals. (*TJ*, p. 473/414)

Moreover, the psychological law that governs this stage of development does not refer specifically to the acquisition of principle-dependent, rather than ideal-dependent, desires (*TJ*, pp. 490–91/429–30). Instead, the acquisition of principle-dependent desires is treated as the upshot of the law (see *TJ*, p. 473/415). And *TJ*'s account of moral development, like that I have found in the revised *Deweys*, depends upon the WOS's institutions "being publicly known to be just by all." I therefore believe that the treatment of moral development in the revised *Deweys* can be read as supplementing, elaborating and clarifying—rather than as fundamentally altering—*TJ*'s statements of the psychological laws governing moral development and its assumptions about the educative effects of publicity. Indeed, though I cannot lay out the argument here, I believe that most of what Rawls says about the acquisition of a sense of justice in *TJ* is compatible with the argument I have extracted from the revised *Deweys*. Even so, Rawls should explicitly have signaled that his thinking about the acquisition of a sense of justice underwent so profound a development. Yet even very late, he maintained that he "would not change that account substantially".[22]

§IX.4: C_3' and the Social Ideals of Justice as Fairness

To show how the Rawls of *PL* would have argued for

> C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and society included in justice as fairness.

21. I am grateful to David Solomon and Daniel Brudney for pressing this question.
22. Rawls, *Restatement*, p. 196, note 17.

it is not enough to show that he thinks members of the WOS would want to live up to an ideal of citizenship. I need to show that he thought they would want to live up to other ideals as well. One of these is a political ideal of friendship. This is an ideal of friendship realized in political life, an ideal of civic friendship. The Rawls of *PL* did think members of the WOS would want to live up to that ideal, but I shall not devote much attention to showing that since the inclusion of this ideal does not constitute a change from his earlier view.

There is one significant difference between his earlier and later *treatments* of the ideal, which I shall mention now and explain later. In *TJ*, Rawls conveyed the impression that the ideal was arrived at by beginning with the concept of a relationship among citizens and specifying it using principles of justice. I believe Rawls has no reason to disavow the earlier treatment. But in a much later essay, "Idea of Public Reason Revisited," he implies that the ideal of civic friendship is arrived at by starting with the idea or the concept of the relation among citizens and using the principles of public reason to specify an ideal of how citizens are to treat one another in public political discussion.[23] Thus, as with the ideal of citizenship, so with the ideal of civic friendship, the political ideal of justice as fairness is arrived at using two sets of principles of right: the principles of justice and "precepts governing reasonable political discussion." Since the sense of justice includes a desire to live up to that ideal, it includes a desire to act from both sets of principles.

What of the third ideal to which C_3' refers? Does the Rawls of *PL* think that the sense of justice includes an ideal-dependent desire that has a social ideal as its object?

We have seen that C_3 refers to an ideal realized by the WOS itself, the ideal of a social union of social unions. We have also seen why Rawls came to think that he could not count on convergence on that ideal. For the Rawls of *TJ* thought that members of the WOS would value participation in a social union of social unions because of a psychological assertion I called the "qualified version of the Companion Effect":

> (4.5') "When men are secure in the enjoyment of the exercise of their own powers, they are disposed to enjoy the perfections of others, especially when their several excellences have an agreed place in a form of life the aims of which all accept" (*TJ*, p. 523/459).

In §§VIII.3 and VIII.5, I argued that Rawls lost confidence in (4.5'). Because he needed (4.5') to get to C_3, this loss of confidence was one of the reasons Rawls came to think C_3 was unrealistic.

On my reading, the Rawls of *PL* began to emphasize a different and less-demanding ideal than the social union of social unions as part of his new

23. Thus at "Public Reason Revisited," *Collected Papers*, p. 579, Rawls says "To make more explicit the role of the criterion of reciprocity as expressed in public reason, note that its role is to specify the nature of the political relation in a constitutional democratic regime as one of civic friendship."

account of stability. Rawls never acknowledged as openly as he should have that his later account did not appeal to the earlier ideal. The closest he ever came to acknowledging the shift was in an important footnote in which he says that he "rebuil[t]" the arguments that appeal to the earlier ideal because "the conception of a social union of social unions . . . is no longer viable as a political ideal once we recognize the fact of reasonable pluralism" (*PL*, p. 388, note 21). Elsewhere, however, he seemed to suggest that the ideal of a social union of social unions remained a part of his account of stability even after the political turn.[24] For reasons I gave in Chapter VIII, I do not believe that the ideal as understood in *TJ* could do any important work in *PL*, and I give the remark about "rebuilding" arguments that appeal to the ideal of a social union of social unions a great deal of weight. Identifying the ideal that took its place helps to confirm that the Rawls of *PL* relied on C_3'.

The natural place to look for a social ideal is "Priority of the Right and Ideas of the Good," where Rawls cites a number of ways in which life in a WOS is good for citizens singly and collectively (*PL*, pp. 201ff). I have already noted in §VIII.3 that that essay omits any mention of a social union of social unions. I now want to zero in on just one of the goods it does mention: "successfully conducting reasonably just . . . democratic institutions over a long period of time." This, Rawls says, "is a great social good and appreciated as such" (*PL*, p. 204).

What is this good? I suggest that citizens conduct just democratic institutions successfully when they support institutions that implement and are known to implement mutually acceptable principles of justice. "Support" can include putting just institutions in place, defending them, and taking part in them. As we shall see in Chapter X when I discuss the liberal principle of legitimacy, it also includes abiding by political outcomes that result from, and that all can recognize as resulting from, proper exercises of various kinds of political power, such as legislative, judicial, executive, and electoral power. I assume that just institutions are democratic institutions and that political power is exercised properly when it is exercised democratically. So when citizens support just institutions in all these ways, they live up to a certain conception—a certain ideal—of democratic politics: "the ideal of citizens governing themselves in ways that each thinks the others might reasonably be expected to accept" (*PL*, p. 218; cf. *PL*, pp. 139–40). Let us call this ideal the *Ideal of Democratic Governance*.

In §IX.3, we saw how Rawls uses principles of right to move from the concept of the citizen, as found in the tradition of democratic thought, to the conception or ideal of citizenship that is found in his later work. The *Ideal of Democratic Governance* results from specifying another of the "basic intuitive ideas"[25] on which justice as fairness is founded: that of society as a fair scheme of social cooperation. That basic idea, like the others, is specified into the

24. Rawls, *Restatement*, p. 142.

25. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 396.

political ideal by reference to principles of right—in this case, the principles of justice, the guidelines of public reason and, as we shall see in Chapter X, the principle of legitimacy. An ideal-dependent desire to live up to the *Ideal of Democratic Governance* would therefore include, or entail the presence of, desires to act from these principles. It would include, or would entail the presence of, dispositions to support just institutions, to debate and vote according to the guidelines of public reason, to regard only legitimate exercises of power as justified, and to acknowledge legitimate exercises as such.

Thus I take it that the *Ideal of Democratic Governance* is one of the ideals to which C_3' refers, and that the desire to live up to that ideal is central to the sense of justice as Rawls understood it in his late work. Because Rawls thinks members of the WOS would acquire a sense of justice, we would expect him to say that members of the WOS acquire the desire to live up to the *Ideal of Democratic Governance*. He does seem to say that, if not as clearly as we might wish. He speaks of citizens' "conception-dependent [i.e. ideal-dependent] desire to have a shared political life on terms acceptable to others as free and equal" (*PL*, p. 98). The desire he refers to in this passage certainly seems like a desire to live up to "the ideal of citizens governing themselves in ways that each thinks the others might reasonably be expected to accept."

Because the desire to live up to the *Ideal of Democratic Governance* is part of citizens' sense of justice, the presence of that desire helps to stabilize justice as fairness. The *Ideal of Democratic Governance* is similar to the ideal of a social union of social unions, since citizens who realize the *Ideal of Democratic Governance* value "the successful carrying out of just institutions" as a final end, just as do citizens who take part in a social union of social unions (*TJ*, p. 527/462). The later ideal is weaker than the earlier one, however, because members of the WOS of *PL* do not "have the common aim of cooperating together to realize their own and []others' nature" in the great diversity of activities a just liberal society permits (*TJ*, p. 527/462). I believe that the *Ideal of Democratic Governance* takes the place of the ideal of a social union of social unions because the Rawls of *PL* realized that the stronger ideal could not do the work he had assigned it in *TJ*. I therefore take it that the *Ideal of Democratic Governance* is introduced as part of the effort to "rebuild[]" the social union of social unions argument (*PL*, p. 388, note 21). Thus on my reading, the ideal—like so much of the conceptual apparatus added to justice as fairness between *TJ* and *PL*—is introduced to help remedy the shortcomings Rawls found in his original treatment of stability.

§IX.5: Whither Congruence?

At the beginning of §IX.1, I said that the Rawls of *PL*, like the Rawls of *TJ*, says that "stability involves two questions," the first of which is the question of whether members of the WOS would acquire a sense of justice (*PL*, p. 140). The argument of §IX.2 shows that the Rawls of *PL* placed ideal-dependent

desires at the center of a sense of justice. The arguments of §§IX.3 and IX.4 show how he thought members of the WOS would acquire a sense of justice so understood. We saw that according to the Rawls of *TJ*, the second question stability involves is whether members of the WOS would judge that their sense of justice belongs to their good "when they assess their situation independent of the constraints of justice" (*TJ*, p. 399/350). This, we saw, is the question of whether the right and the good are congruent.

I have said that on my reading, the Rawls of *PL* thought that stability involves a similar question. It involves asking whether members of the WOS would judge it rational, in light of their comprehensive views of the good, to maintain and act on their sense of justice—understood now as the ideal-dependent desires discussed in the previous sections. That question is answered in the affirmative by the conclusion of what I called *PL*'s "basic stability argument":

> C_9: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

As I have already noted, my reading seems to face at least two serious textual difficulties. One is that the Rawls of *PL* never describes the second question as a question of congruence, as we would expect him to do if the two treatments of stability run parallel. The other is that the second question I have said stability involves is not the second question the Rawls of *PL* says it involves. In the version of "Idea of an Overlapping Consensus" included in *PL*, he writes:

> The second question [stability involves] is whether in view of the general facts that characterize a democracy's public political culture, and in particular the fact of reasonable pluralism, the political conception can be the focus of an overlapping consensus. (*PL*, p. 141)

This passage seems to pose a serious challenge for my reading. I described the "basic stability argument" that I laid out in §IX.1 as *basic* because it supports an affirmative answer to what I identified as the second stability question of *PL*. We saw that that argument is premised on the possibility of an overlapping consensus at the second and third steps. If the second question involved in stability is the question Rawls poses in the passage I just quoted, then we would expect that (9.2) and (9.3) would express the *conclusion* of *PL*'s stability argument, not two of its *premises*. So the passage suggests that I have mistaken how the Rawls of *PL* argues for stability.

I believe, on the contrary, that my reading is faithful to Rawls's thought. Rawls insists repeatedly that an overlapping consensus is a *condition* of stability (e.g., *PL*, p. 44), thereby suggesting that stability can be shown if we suppose—as I have in laying out the basic stability argument—that that condition is satisfied. Moreover, the Rawls of *PL* professes a concern to show how "the existence and public knowledge of a reasonable overlapping consensus" stabilize a WOS (*PL*, p. 392)—a concern that, as we shall see in §X.5, is

especially evident in the "Reply to Habermas." The basic stability argument has the merit of showing how that concern is answered. Furthermore, I grant that Rawls's argument for an affirmative answer to the second question I have said stability involves depends critically upon (9.2) and (9.3). It therefore depends critically upon answering the second question Rawls says stability involves. Since so much rests on those steps in the basic stability argument, we can see why Rawls might have cut to the heart of the matter and posed *it* rather than the question I have taken him to be trying to answer.

Thus, two of the worries about my reading that are raised by the quoted passage can, I believe, be explained away. But if the Rawls of *PL* did indeed divide the stability problem in the way that I have said, why didn't he ever repeat *TJ*'s claim that stability "involves" the question of whether the right and the good are congruent?

Some readers think that an argument that an overlapping consensus obtains or is possible *is* at the heart of a new argument for congruence. For, these readers will insist, when a conception of justice is the focus of an overlapping consensus, the public conception of right is "supported" by reasonable conceptions of the good (*PL*, p. 145). And when this support relation obtains, the congruence of the right and the good follows.[26] On this reading, Rawls continued to think that showing stability requires him to show congruence; an overlapping consensus furnishes a different way of showing it. But this reading is mistaken, for to say that justice and goodness are congruent is not just to say that it is good to be just or that citizens' views of the good somehow support their sense of justice. Congruence is an especially strong support relation. By the time he wrote *PL*, Rawls recognized that a WOS could be stable without it.

We saw in §II.3 that congruence is a relation between two points of view within practical reason: the OP and the point of view of full deliberative rationality. It obtains when members of the WOS, reasoning within the latter point of view, conclude that their balances of reasons tilt in favor of treating the desire to act from principles chosen in the OP as regulative of their plans of life. That is why, as we have seen, the Rawls of *TJ* tried to show congruence by producing various arguments that eventually lead to a conclusion about the viewpoint of full deliberative rationality, the conclusion I called the *Congruence Conclusion*:

> C_C: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

When I introduced congruence, I suggested that establishing congruence establishes a relationship of subordination between one point of view within

26. See Paul Bou-Habib, *Review of The Cambridge Companion to Rawls* (Cambridge: Cambridge, 2004), ed. Freeman, *Journal of Moral Philosophy* 1, 3 (2004): pp. 375–79, p. 377.

practical reason and the other, and that the Rawls of *TJ* thinks this relation *unifies* our practical reason. In §VII.5, I tried to indicate what I think Rawls has in mind. Our scheduling and pursuit of various ends are not properly unified by subordination of all other ends to some one end which is dominant. Rather, each person's reason is unified when she makes and executes plans that are regulated by principles of right. That is part of why each person would judge that maintaining her sense of justice as supremely regulative is itself good for her, even when judged from within the thin theory. That judgment, and common knowledge of it, help to stabilize justice as fairness.

As we have seen, the problem with *TJ*'s treatment of stability was that it established congruence by way of *TJ's Nash Claim*:

C_N: Each member of the WOS judges, from within the thin theory of the good, that her balance of reasons tilts in favor of maintaining her sense of justice as a highest-order regulative desire in her rational plans when the plans of others are similarly regulated.

The arguments for C_N failed. They failed because they depended, in part, upon an account of the unity of reason that was supposed to hold of everyone and upon a form of unity that each person was supposed to maintain because she judged it to be good. But as we saw in Chapter VIII, not everyone accepts a Kantian account of the unity of the self and not everyone thinks that taking the desire to act from the principles as regulative of her plan of life in its entirety is good for her. In a pluralistic society, the stability of a conception of justice cannot rest upon a general account of how reason is unified or of how goodness and justice are related.

The Rawls of *PL* recognized this. His recognition is shown in the basic stability argument. If the Rawls of *PL* relied on the existence or the possibility of an overlapping consensus to show congruence and stability, we would expect his argument for stability to move from an overlapping consensus—asserted at (9.2) and (9.3)—to *PL's Nash Claim* and C_9, and from there to stability via the *Congruence Conclusion* C_C. But that is not how the argument goes. Rawls concludes that the WOS would be stably just on the basis of C_{PL}, and C_{PL} is a weaker claim than the *Congruence Conclusion.*

C_{PL} says:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

To see that this claim is weaker than the *Congruence Conclusion*, recall what is meant by saying that the sense of justice is treated as "a highest-order regulative desire in their rational plans." When congruence obtains, each person treats the principles of justice as the most fundamental or—as I put it earlier—the *ultimate* ethical principles in her rational plan, provided others do as well. The results given by the principles of justice are not "checked" against the

deliverances of other ethical principles, such as those derived from natural law or from religious sources. This is why the *Congruence Conclusion* is so strong and why, as we saw in §VIII.4, Rawls ran into difficulty with showing it. We shall see that if the WOS is to be stably just, then members of the WOS must take their sense of justice as regulative of their political lives. But C_{PL} can be true while they take *other* ethical principles to be ultimate—principles that imply or in some other way support the principles of justice. It is also possible that some members of the WOS maintain the principles of justice, but do not take *any* principles to regulate their plans in their entirety because the components of their plans are relatively independent. C_{PL} is therefore weaker than congruence, but strong enough for the Rawls of *PL*.

Moreover, on my reading, Rawls does not need to show congruence to establish either *PL's Nash Claim* or C_9, the claim that each member of the WOS maintains justice as fairness on the basis of her comprehensive view. Instead, I have said that he can get to them from the weaker:

(9.3) Each comprehensive doctrine is "*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 for a democratic regime" (*PL*, p. 169, emphases added).

Rawls said in *TJ* that "the nature of the self as a free and equal moral person is the same for all, and the similarity in the basic form of rational plans expresses this fact" (*TJ*, p. 565/495). In §VII.5, we saw what the Rawls of *TJ* meant by that. Between *TJ* and *PL*, Rawls came to appreciate that in a liberal society, there are many forms people's lives and plans can assume. Those who hold liberal comprehensive doctrines may find it easy to unify those doctrines with their political views. Citizens of traditional faith, however, may experience profound tension between their faith and their political views, a tension that is only resolved with difficulty, if at all. For these members of the WOS, it is a daily struggle to—as Richard Rorty once said—"hold justice and reality in a single vision."[27] No one account will explain how, if at all, all the members of the WOS achieve this singularity of vision. And so by *PL*, Rawls concluded that while the nature of the *citizen* as free and equal is the same for all, "it is left to citizens individually…to settle how they think the values of the political domain are related to other values in their comprehensive doctrine" (*PL*, p. 140). Thus, it is only in *PL* that the WOS realizes one of the most important promises of a liberal society. That promise is hinted at in *TJ*, where Rawls said that "the parties' aim in the original position is to establish just and fair conditions *for each to fashion his own unity*" (*TJ*, p. 563/493, emphasis added).

27. Richard Rorty, "Trotsky and the Wild Orchids," in his *Philosophy and Social Hope* (New York: Penguin Press, 1999), pp. 3–20, p. 7.

X

Comprehensive Reasons to Be Just

In Chapter IX, I began to examine the reconstructed account of stability Rawls developed in *Political Liberalism*. In §IX.1, we saw that in *PL* as in *TJ*, "stability involves two questions" (*PL*, p. 141). We also saw that in *PL* as in *TJ*, the first of these questions is whether members of the well-ordered society (WOS) would acquire a sense of justice. In §IX.2, I showed how Rawls changed his description of a sense of justice between *TJ* and *PL* so that conception- or ideal-dependent desires are central to it. In §§IX.3 and IX.4, I showed that he argues members of the WOS would normally acquire a sense of justice by arguing for:

> C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and society included in justice as fairness.

In *PL* as in *TJ*, showing stability also requires showing that members of the WOS would all judge, from the viewpoint of full deliberative rationality, that maintaining the desire to live up to the ideals of justice as fairness belongs to their good. And in *PL* as in *TJ*, what is most interesting about the second part of the stability argument is Rawls's attempt to show that members of the WOS would think that it is good to be just even when they step outside the viewpoint of full deliberative rationality and leave out of account their desires to be just for its own sake. Reaching this judgment requires them to adopt a somewhat artificial perspective on their desires. In *TJ*, the perspective is that of someone "following the thin theory of the good" (*TJ*, pp. 569–70/498). In *PL*, it is the perspective of persons judging "by their comprehensive view[s]" (*PL*, p. 392). What I called *PL*'s "basic stability argument" purports to show that members of the WOS would judge that it is good to be just when they adopt that perspective.

In §IX.1, I said that *PL*'s basic stability argument goes as follows. Rawls begins by assuming that:

(9.1) Members of the WOS follow their comprehensive doctrines.

He thinks that if an overlapping consensus obtains:

(9.2) "Reasonable doctrines endorse the political conception, each from its own point of view" (*PL*, p. 134).

We saw that Rawls elaborates what he means by "endorse." If an overlapping consensus on a conception of justice obtains then:

(9.3) Each comprehensive doctrine is "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 for a democratic regime" (*PL*, p. 169).

(9.3) implies that:

(9.4) Each comprehensive doctrine is "either congruent with, or supportive of, or else not in conflict with" the political ideals referred to by C_3' and the values of realizing them.

When any one of the relations "congruent with," "supportive of," and "not in conflict with" holds, then:

(9.5) According to each comprehensive doctrine, the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them" (*PL*, p. 156).

(9.4) and (9.5) imply:

(9.6) Each member of the WOS judges, from within her comprehensive view, that the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them," at least when others reach the same judgment.

From (9.6), Rawls infers *PL's Nash Claim*:

C_N*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

In §IX.5, we saw that the basic stability argument of *PL* is *not* an argument for congruence. In this chapter, I want to look closely at the assumption that an overlapping consensus obtains in a WOS. Examining this assumption will show why Rawls introduces the notion of political legitimacy in *PL*—a matter which is, I believe, much misunderstood. As I implied when I introduced the basic stability argument in §IX.1, it also shows how the Rawls of *PL* solves the *mutual assurance problem*. Solving that problem allows him to reach:

C_9: "citizens will judge (by their comprehensive view) that political values either outweigh or are normally (though not always) ordered prior to whatever nonpolitical values may conflict with them" (*PL*, p. 392).

From C_9, it is a short step to the conclusion about full deliberative rationality that Rawls really wants, a conclusion I expressed as:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

Finally, I shall look at why Rawls came to think that a WOS would be characterized by an overlapping consensus on a family of reasonable political conceptions rather than on justice as fairness alone.

§X.1: Moving from (9.2) and (9.3) to (9.5)

Early in *PL*, Rawls says that the fundamental question about political justice in a democratic society 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*, pp. xxxix, 4). Believing—for reasons I shall take up in §X.2—that citizens of faith pose this question in an especially interesting and pressing form, and that "stability" has to be understood as "stability for the right reasons," Rawls recasts the questions as "How is it possible for citizens of faith to be wholehearted members of a democratic society when they endorse an institutional structure satisfying a liberal political conception of justice with its own intrinsic political ideals and values, and when they are not simply going along with it in view of the balance of political and social forces?" (*PL*, p. xl).

The first formulation of the question refers to a "stable society," by which we know—from §II.1—Rawls means a "stably just society" rather than a society that enjoys what I called "state stability." The second reformulation concerns the possibility of citizens' "endorsing" an "institutional structure." This shift of terminology does not betray a shift in Rawls's concerns. Rather, once he recognized—as he did in *PL*—that members of a WOS might disagree about which conception of justice was correct, it was natural for him to express his concern about a stably just society by asking whether citizens could agree on structure of government which they recognize as reasonably just. The phrasing of the question therefore reflects a complication which I have said I shall ignore until late in this chapter for simplicity's sake. The argument that citizens of the WOS would all acquire a sense of justice goes some way toward answering that question, since those with a sense of justice support the just institutions under which they live. What the argument does *not* show is that they will support it "*wholehearted[ly]*." Why is it important to show wholeheartedness of support?

As we saw in §II.3, the Rawls of *TJ* was deeply concerned with whether justice as fairness would be destabilized because practical reason is at odds with itself. His arguments for the congruence of the right and the good were supposed to help show stability by showing that the two powers of practical reasoning, which he later called the Reasonable and the Rational, would not pull citizens of the WOS in different directions. Rawls summed up this line of thought much later in an important passage that I have quoted before, saying the institutions [of the WOS] "are stable because the just and the good are congruent. That is, no reasonable and rational person in the well-ordered society of justice as fairness is moved by rational considerations of the good not to honor what justice requires."[1] On one interpretation, showing that they can support just institutions wholeheartedly removes this source of instability by showing that they would not be divided. The way Rawls reformulates the fundamental question of *PL* as a question about the possibility of wholehearted support therefore shows another continuity between the *TJ* and *PL*, for it shows Rawls's continued concern that citizens of the WOS be able somehow to unify their powers of practical reasoning.

TJ's congruence arguments foundered on the possibility that citizens of the WOS *would* be pulled in different directions by their sense of justice and their comprehensive conceptions of the good, and so *would* "be moved by rational considerations of the good not to honor what justice requires," at least in some cases. *PL*'s basic argument for stability—the argument for $C_N{}^*$, C_9, and C_{PL}—shows how wholehearted support is possible. For it shows that members of the WOS, including citizens of faith like one I shall call Jan, affirm their sense of justice from within their comprehensive doctrines. But as I said at the close of Chapter IX, in *PL* unlike *TJ*, there is no one set of arguments about how practical reason is to be unified or wholeness of heart is to be achieved. Rather, how citizens unify their conceptions of the good and their sense of justice is left up to them.

As we saw in Chapter IX, the assumption of an overlapping consensus is one of the critical moves in the basic stability argument. Without that assumption, the argument for stability would seem to depend—implausibly—either upon knowledge of each member of the WOS singly or upon there being some one support relation between justice as fairness and all reasonable comprehensive doctrines. With the assumption, and the move from the assumption to (9.5), the argument for *PL's Nash Claim* $C_N{}^*$, at least, seems straightforward. Once Rawls solves the *mutual assurance problem*, he can reach C_9 and C_{PL} as well.

The existence of an overlapping consensus is assumed at steps (9.2) and (9.3). One obvious question about the argument concerns its seeming equation of these two steps. (9.3) leaves open the possibility that some comprehensive doctrine is "not in conflict with" the values specified by the political conception

1. Rawls, "Domain of the Political," *Collected Papers*, p. 487, note 30.

of justice. This seems to make (9.3) considerably weaker than (9.2), which says that comprehensive doctrines all "endorse" the political conception. Of course, if (9.3) is strong enough to support the move from (9.4) to (9.5), then (9.3) is much stronger than it appears and the equation of (9.3) with (9.2) may not be so implausible. But the wording of (9.3) makes it hard to see how (9.3) could be strong enough for that. So the real difficulty with the argument, and the one I shall take up in this section, seems to be that of getting from (9.3), via (9.4), to (9.5). In the next section, I shall ask whether Rawls is justified in assuming that an overlapping consensus would obtain.

In "Idea of an Overlapping Consensus," Rawls considers a "model case" of an overlapping consensus in which the political conception of justice is supported by four comprehensive doctrines: value pluralism, Kantianism, utilitarianism, and a religious view that includes a doctrine of free faith (see *PL*, p. 169). He takes up these four doctrines because together they exemplify the three relations in which (9.3) says reasonable comprehensive doctrines stand to the values of a political conception of justice when an overlapping consensus obtains. Rawls assumes that these three relations exhaust the relations that can obtain between a political conception and reasonable doctrines when an overlapping consensus obtains; later we shall see that this depends upon the looseness of "supportive of." For now, it will be useful to look at each case, see in what way (9.3) is true of it, and see whether Rawls can move to the relevant instance of (9.5).

Let me begin with value pluralism. Value pluralists think that "each domain of value has . . . its own free-standing account" (PL, p. 170). So according to comprehensive value pluralism, there are different domains of human life, such as the political, the ethical, and the aesthetic. The values to be realized in each domain of life are, like the political values of justice as fairness, domain-specific. These domain-specific values may be amenable to systematic accounts, and, as with justice as fairness so with all the others, no one of the accounts presupposes the conclusions of others. That is what makes them free-standing. The conjunction of the free-standing accounts is the comprehensive view.

When an overlapping consensus obtains, the account of the political domain that value pluralism includes just is justice as fairness. In this case, as Rawls says, value pluralism "include[s] the political conception as the part covering political values" (*PL*, p. 170). Since value pluralism is a conjunction of free-standing accounts without any overarching philosophical unity, conflicts among the various components are settled by balancing (*PL*, p. 145). If there is a prima facie conflict between ethical and political values, Rawls says "in this comprehensive pluralist view the political conception is affirmed by balancing judgments that support the great values of the political against whatever values normally conflict with them" (*PL*, p. 170). It then follows immediately that according to value pluralism, the values and ideals of justice as fairness "normally outweigh whatever values oppose them" (*PL*, p. 155) and so the relevant instance of (9.5) will be true of it. Adherents of the view will then judge that those ideals and values have priority, just as (9.6) implies.

I just supposed that there are prima facie conflicts within value pluralism, and that they will be settled by balancing values as the value pluralist thinks best. But (9.3)'s implication that some comprehensive views are "not in conflict with" justice as fairness suggests another possibility. It suggests the possibility that unlike the ethical components of Kantianism, utilitarianism, and the religious view, the ethical component of value pluralism has no political implications. Since its ethical component has no political implications, value pluralism that includes the political conception does not recognize any values that compete with political values, and (9.5) is trivially true of value pluralism.

It is tempting to think of value pluralism as the doctrine so famously championed by Isaiah Berlin, and to think that it is held only by philosophical sophisticates. But I think Rawls believes that most members of the WOS are really value pluralists of some kind. This is suggested by his remark that most comprehensive doctrines "are not seen by [their adherents] as fully general and comprehensive" (*PL*, p. 160). Rather, as held, the religious and ethical components of these doctrines do not have "any particular connection, one way or the other" with the political conception (*PL*, p. 160). There can be, Rawls says, "a certain looseness" in comprehensive views (*PL*, p. 159) that "allows scope for the development of an independent allegiance to the political conception" (*PL*, p. 168). Citizens who profess to be Kantian or religious or Green, for example, may really endorse Kantianism as a form of personal morality, or religion or deep ecology as forms of devotion. But their Kantianism or their religion or their reverence for the earth may be restricted to the domains of the ethical or the aesthetic or the spiritual and they may find that, even on reflection, they do not endorse purported social implications of those views. When it comes to politics, they just accept the political conception as the right account.

Clearly, these citizens' acceptance of justice as fairness is facilitated by the re-presentation of justice as fairness as a political conception of justice. But some members of the WOS may *not* be value pluralists. Instead, they endorse comprehensive views, the ethical components of which have some social and political implications.

One example of such a view, which Rawls considers in the model case, is Kant's view or "a view sufficiently similar to it" (*PL*, p. 169). I believe what Rawls has in mind is that those who accept these views accept full autonomy as an ethical ideal to be realized in the whole of life, including political life, and they believe they can unify themselves only by treating the desire to act from principles of right as supremely regulative. The part of their Kantianism that bears on political life is not identical with "the publicly recognized political conception of justice" precisely because the latter is a political conception. Rather, their Kantianism is a comprehensive liberalism. The question, then, is what relation they see between their Kantianism and the political conception.

The answer, I think, is that according to this form of comprehensive liberalism, adherents recognize that the best way to live up to its ideals in political

life is to live up to the political ideals of justice as fairness and to take its principles as finally authoritative. It is clear enough why Rawls would say this, since he presumably wanted to leave open—and to consider—the possibility that some members of the WOS might endorse the Kantianism of his own earlier work. It certainly seems plausible that the best way to live up to the ideals of Rawls's early Kantianism, at least in political life, is to live up to the ideals of the political liberalism expressed in his later work.

This case is, I believe, the one Rawls has in mind when he implies in (9.3) and (9.4) that some comprehensive views held in the WOS are "congruent" with the values of justice as fairness. Whether or not this form of Kantianism provides a "deductive basis" (*PL*, p. 169) for the political conception, as Rawls seems to imply, it surely ranks the ideals of that conception above potentially competing values. So as in the case of comprehensive doctrines that are "not in conflict" with justice as fairness, so in the case of those that are congruent with it, the relevant instance of (9.5)—and hence of (9.6)—holds. Kantians in the WOS have a sense of justice informed by justice as fairness, and so want to live up to the ideals of justice as fairness. When they ask themselves whether they should include the satisfaction of those desires in their plan of life, they assume that the only reasons they have to do so are provided by their Kantianism. But since (9.6) holds of them, they take themselves to have overriding reasons to satisfy those desires. So *PL's Nash Claim* is true of them as well; if we assume a solution to the *mutual assurance problem*, then so are C_9 and C_{PL}.

I believe Rawls thinks that if an overlapping consensus holds, then other comprehensive doctrines with political implications, such as utilitarianism and religion with a doctrine of free faith, will exemplify the third relation referred to in (9.3): they will be "supportive of" justice as fairness. What Rawls has in mind, I think, is that adherents of such views—like Kantians—will think that the best way to live up to their ideals and values in the political life of the WOS will be to live up to the ideals and values of justice as fairness. In the case of these doctrines, the claims about how best to live up to them will not depend upon their "congruence" with justice as fairness, as in the case of Kantianism. It may depend, instead, upon claims about the best "workable approximation" of a comprehensive view in political life (*PL*, p. 170). But having established the claim he wants about how best to live up to the ideals of these views, Rawls can—despite the difference in how he arrived at that claim—then say that these views "support" justice as fairness. He can then move from the relevant instance of (9.3) to the relevant instances of (9.5), (9.6) and *PL's Nash Claim*, C_9 and C_{PL} just as he did in the case of Kantianism.

This discussion of the model case, and its implications for the move from (9.3) to (9.6) that I just sketched, fits with my assertion of an interesting parallel between Rawls's earlier and later treatments of stability. Since C_3' is true, everyone in the WOS has acquired a sense of justice informed by justice as fairness—understood, as we saw in §IX.3, as desires to live up to the ideals of justice as fairness. In *TJ*, a crucial step in establishing that members of the

WOS would judge that justice belongs to their good is the supposition that the typical member of the WOS gives no independent weight to her sense of justice as such and "follow[s] the thin theory of the good" (*TJ*, pp. 569–70/499). In *PL*, a crucial step in reaching that conclusion is the supposition that members of the WOS do not give independent weight to their sense of justice as such. Instead, they assume that the only reasons they have to be just are *comprehensive reasons*: reasons provided by their comprehensive views of the good. For purposes of showing stability, value pluralists assume that the only reasons they have to satisfy the ideal-dependent desires to which C_3' refers are reasons provided by their pluralism. Comprehensive Kantians assume that the only reasons they have to satisfy those desires are the reasons provided by their Kantianism. Comprehensive utilitarians and adherents of religions endorsing free faith assume that the only reasons they have to satisfy them are provided by their utilitarianism and their religion. But since (9.6) is true of them, they judge—from within their pluralism, their Kantianism, their utilitiarianism, or their faith—that the ideals and values of justice as fairness have overriding weight in political life. So those who endorse these comprehensive views will conclude—again, from within their views—that the reasons they have to live up to those ideals and values outweigh reasons to pursue competing values, and *PL's Nash Claim* will be true of them and, once the *mutual assurance problem* is solved, C_9 and C_{PL} are shown true of them as well.

§X.2: Would there Be an Overlapping Consensus?

Thinking through the model case shows that (9.3) is strong enough to bear the weight that Rawls puts on it. It also shows why Rawls seems to equate (9.3) with (9.2). Views of the good that are "congruent with" or "supportive of" a political conception of justice clearly endorse it; value pluralism, the view which is not in conflict with it, endorses it as well. And so the case shows how an overlapping consensus stabilizes a WOS: *if* an overlapping consensus obtains, then everyone has a sense of justice and affirms it as part of her good. But why think—as the basic stability argument assumes between the first and second steps—that an overlapping consensus would obtain in a WOS? In particular, why think that conceptions of the good which have political implications—all conceptions that are not pluralist, and some pluralist views as well—would endorse justice as fairness, as (9.2) says, or would be "either congruent with, or supportive of" justice as fairness, as (9.3) asserts?

Religious comprehensive views may not be the only comprehensive views that raise this question, but they can exemplify the problems of moving toward an overlapping consensus in useful ways. I shall therefore consider how religious views might come to be included in one.

The constitution and the laws of the WOS implement the principles of justice. Since the first principle requires equality of the basic liberties, I assume that the constitution and the laws of the WOS would be religiously neutral. They

would not grant any church preferred status, and citizens would be free to choose, change, or reject religion as they like, without state dissuasion or encouragement. So if a religious citizen of the WOS, Jan, has a sense of justice, she will support religiously neutral institutions. Justice as fairness will be "as stable as one can hope for" (*TJ*, p. 399/350) if she also affirms her sense of justice—including her support for a religiously neutral constitution—from within her comprehensive view.

Suppose, though, that Jan's religion teaches that it offers the surest way to redemption and that people are therefore most likely to be saved if they accept it. Suppose her religion also teaches that they are most likely to accept it if it is given a privileged place in national life by the constitution and is encouraged by legislation that fosters religious culture, allows religious discrimination, favors the religious schools run by Jan's denomination, and imposes liabilities on schools and churches operated by other religions. It is the job of government to promote the general welfare and to give to each his due. When society is arranged in these ways, it might seem to be doing what it can for the—eternal—welfare of its members, and to be properly acknowledging God.

If this is what Jan's religion says government should do, how can Jan help but regard a religiously neutral society as failing in important ways? More specifically, how can she regard it as justified, and acquire a sense of justice that supports it? The basic stability argument presupposes that all members of the WOS—including Jan—have acquired a sense of justice that moves them to support just institutions, but how can that presupposition be true? Moreover, Jan's faith may seem to entail that a society that violates religious neutrality realizes extremely important goods, goods which we might call the *Goods of Salvific Promotion*. If an overlapping consensus is to obtain, Jan's religion must hold that the ideals and values of justice as fairness "normally outweigh" the *Goods of Salvific Promotion*. But how can it? And if it cannot, how can Jan affirm her sense of justice?

This problem becomes clearer when we recall what "normally outweigh" means. To say that the one good outweighs the other in this context is to make a claim about which set of values is better justified or supported by reasons. The conclusion Jan's religion must reach if it is to take part in an overlapping consensus is the conclusion that, when a constitution or a piece of legislation is to be written, the balance of reasons tilts toward religious neutrality rather than toward the *Goods of Salvific Promotion*.

The teachings of Jan's church about the importance of salvation to individual welfare and the role of the church in promoting it would, if true, seem to provide very weighty reasons in favor of the *Goods of Salvific Promotion*. And so it may seem that Jan's religion could reach take part in an overlapping consensus only if it denied its traditional teachings. But if this were the only way for it to reach the conclusion that laws and constitutions should be religiously neutral, it seems impossible—or, if not impossible, at least unlikely—that Jan's religion would "endorse" justice as fairness and take part in an overlapping consensus. So not only does the case of Jan cast a presupposition

of the basic stability argument in doubt—namely, the presupposition that Jan has a sense of justice informed by justice as fairness—but it is also hard to see how two of the steps of that argument—(9.2) and (9.5)—can be true.

Rawls has two answers to this problem. I shall explore the first in the remainder of this section; I shall take up the other, which I believe to be of somewhat greater philosophical interest, beginning in §X.3.

One answer is that comprehensive views can and do *change* so that they can come to take part in an overlapping consensus. Recall that Rawls assumes members of the WOS are reasonable. They want to cooperate with others on terms that are mutually justifiable (*PL*, p. 49). This desire is part of their moral sensibility, and is itself encouraged by the just institutions under which they live.[2] Because they are reasonable, they affirm only what Rawls calls "reasonable comprehensive doctrines" (*PL*, p. 59). Rawls's definition of a reasonable comprehensive doctrine is quite weak (*PL*, p. 59). The third feature is especially important in the present connection. Reasonable comprehensive doctrines are "not necessarily fixed and unchanging"; they can change "in light of what, from [their] point of view, [they] see as good and sufficient reasons" (*PL*, p. 59). This amenability to change means that even doctrines that are historically anti-liberal can develop under the lived experience of liberal institutions, so that they can join an overlapping consensus on a liberal political conception of justice.

An example of what Rawls has in mind is this. Consider some adherents of Jan's religion who are reasonable persons in Rawls's sense. Because they are reasonable, they will want to repudiate intolerant strains in their tradition. And yet they may also have a great stake in maintaining that they are faithful adherents of their religion. It may be very important to them to think of themselves as orthodox and to assert their own orthodoxy. This may be important to them even if they hold positions that vary from what has traditionally been regarded as orthodox because they have ideal-dependent desires to live up to the demands of justice as fairness. Rather than dissenting, they may therefore offer arguments to the effect that their politics is faithful to their tradition. Over time, other adherents of their doctrine—including those responsible for defining doctrinal orthodoxy, if such there be—may come to appreciate that political values, such as religious toleration and the *Ideal of Democratic Governance*, are such great values that they must be accommodated within the tradition's intellectual framework and that they outweigh the *Goods of Salvific Promotion*. When this happens, their religion develops, so that doctrinal arguments for a liberal constitution are accepted from within the doctrine. When they are accepted, reasons that once seemed to support the heterodoxy of political liberalism are recognized as "good and sufficient" from the point of view of the doctrine itself. In this way, comprehensive doctrines evolve to embrace liberal constitutional provisions and the values and principles that support them.

2. See the important footnote at *PL*, p. 85.

The evolution of comprehensive doctrines is a long and complex process. Rawls provides a very schematic, hypothetical history of that process in "Idea of an Overlapping Consensus" (*PL*, pp. 158–68). I have skipped even the minimal history that Rawls provides, since my example presupposes that just institutions are already in place. Adequate discussion of even a single case of doctrinal evolution would require considerably more space—and considerably more historical, textual, and theological expertise—than I possess.[3] Here I note just two points about the evolution that Rawls says could help to bring about an overlapping consensus.

First, this evolution depends upon adherents of the comprehensive doctrine developing allegiance to the political conception and its ideals. The development of this allegiance depends, in turn, upon what liberal democratic institutions do and are known to do. In a WOS, these institutions institutionalize and publicize the liberal conception of justice. Thus, doctrinal evolution ultimately depends upon the terms of the cooperation that are implemented. This evolution, when fully laid out, shows how "a reasonable and effective political conception"—by which I take it Rawls means "one which effectively regulates basic institutions and instills a sense of justice"—"may bend comprehensive doctrines toward itself" (*PL*, p. 246).

We saw in earlier chapters that the Rawls of *TJ* argued that justice as fairness would, when institutionalized and publicized, generate its own support by fostering a sense of justice. It would also generate its own support by encouraging members of the WOS to converge on an important set of ideal-dependent desires and on desires that provide thin reasons to be just. These arguments were central to *TJ*'s case for the inherent stability of justice as fairness. The Rawls of *PL* thinks justice as fairness generates its own support in much the same way. When it is institutionalized and publicized, it can foster a sense of justice, understood now as a set of ideal-dependent desires to live up to the ideals and values of justice as fairness, and it can encourage convergence of comprehensive doctrines on those ideals and values, so that (9.2) and (9.3) are true and an overlapping consensus obtains.

But second, though doctrinal evolution may help to bring about an overlapping consensus, evolution of this kind may require more liberalization of comprehensive doctrines than might realistically be expected. This is especially

3. The classic treatment of doctrinal development in the Christian tradition is, of course, John Henry Newman, *An Essay in the Development of Christian Doctrine* (Whitefish, MT: Kessinger Publishing, 2007). Some of the arguments that proved most important in moving the Catholic church toward its embrace of the doctrine of free faith at the Second Vatican Council were those of John Courtney Murray, especially those published in *Theological Studies* from 1948 onward. A complete bibliography of Fr. Murray's writings can be found at: http://woodstock.georgetown.edu/library/Murray/0_murraybib.html. For an extended historical argument that the development of Catholic doctrine on toleration was due to the American Catholic experience, see John T. Noonan, Jr., *The Luster of Our Country: The American Experience of Religious Freedom* (Berkeley, CA: University of California Press, 2000).

clear when we ask, not about how adherents of Jan's religion might regard a religiously neutral constitution, but how they might regard particular pieces of legislation. Suppose Jan's religion teaches that religiously neutral legislation, legislation authorizing the use of military force, the legalization of same-sex marriage, or legislation permitting access to abortion is contrary to natural law and so is not really binding law at all. Her religion *may* liberalize these teachings, but it may not.

As we saw, the basic stability argument presupposes that Jan has acquired a sense of justice informed by justice as fairness. According to that argument, if an overlapping consensus obtains, then Rawls can reach *PL's Nash Claim*, C_N*. And if we assume an answer to the *mutual assurance problem*, then Rawls can get to the remaining conclusions I have said he wants to reach. At the beginning of the previous section, I said that showing how these conclusions can be true of Jan and other citizens of her faith shows how they can be "wholehearted members" of a WOS who "endorse [its] institutional structure . . . with its own intrinsic political ideals and values" (*PL*, p. xl). They need not simply "go[] along with [their society's basic arrangements] in view of the balance of political and social forces" (*PL*, p. xl). Indeed, I take these three conclusions to be what gives meaning to Rawls's talk of "wholeheartedness."

The problem is that Jan can be a wholehearted supporter of her society's "institutional structure" only if she thinks that the operation of the basic structure yields outcomes that are justifiable when the most important matters are at stake. If she does not, then she will not acquire a sense of right that supports that structure—in which case C_N*, C_9, and C_{PL} cannot be true of her, and she is bound to regard her subjection to those outcomes as a mere acquiescence to the balance of political power. But the claim that legislation legalizing abortion or authorizing military force is justified seems to be incompatible with the traditional and unchanging teachings of her church. So if she does acquire a sense of right that supports the basic structure, she will judge that sentiment to be at odds with her true (religious) good. In that case, C_N*, C_9, and C_{PL} cannot be true of her either.

An overlapping consensus thus seems to require that Jan hold incompatible claims about the justifiability of the law. On my reading, Rawls tried to show compatibility, and to show how an overlapping consensus is possible, by appealing to a form of justification that is anticipated in *TJ* but that is only developed in *PL*: legitimacy. In the next section, I shall look at the liberal principle of legitimacy in some detail since the principle, and Rawls's reasons for introducing it, are so frequently misunderstood.

§X.3: Legitimacy and Justification

The liberal principle of legitimacy says:

> 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 light of principles and ideals acceptable to their common human reason. (*PL*, p. 137)

The beginning of the passage suggests that legitimacy attaches in the first instance to exercises of ordinary political power that yield particular political outcomes, such as laws, judicial decisions, and policies. This is a mistake. Late in *PL*, Rawls says of legitimacy that "reasonable citizens understand this idea to apply to the general structure of authority" (*PL*, p. 393); the context of this passage makes clear that Rawls is referring to the general structure of *political authority* or the constitution. As we shall see, Rawls think that "legitimacy" applies in the first instance to the exercise of power to write a constitution. It is because the constitution that results is legitimate that particular exercises of power can be legitimate as well.

Let us understand a constitution broadly as a set of rules and practices, whether written or unwritten, that specify the essential structure of government. They determine how a society's political authority is to be divided (if at all), how it is transferred, and who is responsible for what political decisions. The essentials of the constitution also state the purposes for which political power is exercised and the limits on its exercise.

In *TJ*, Rawls imagines that the principles of justice are chosen in the first stage of what he calls a "four-stage sequence," and that they are applied to the WOS in the subsequent stages. The constitution of a WOS is to be written at the second stage, which we might think of as a constitutional convention. There, participants decide on a constitution that incorporates the two principles. This convention is not a historical event. Rather, the four-stage sequence is a theoretical device. "It sets out a series of points of view from which the different problems of justice are to be settled, each point of view inheriting the constraints adopted at previous stages" (*TJ*, p. 200/176). But how, exactly, are those writing a constitution to apply or incorporate the principles of justice into "the general structure of authority"?

In §II.3, when I introduced the viewpoint of full deliberative rationality, I said that points of view are given, in part, by the rules of reasoning to be followed by those who occupy it. The viewpoint of the constitutional convention must therefore be specified, in part, by principles that guide the reasoning of those who are trying to apply principles of justice to the constitution. Rawls says that "guidelines and criteria" for applying the principles of justice are the liberal principle of legitimacy and the guidelines of public reason (*PL*, p. 225). If these principles are strong enough to provide real guidance, something has to be said about the source of their normative force. In justice as fairness, the most basic normative principles are those chosen for the basic structure in the original position. The principles guiding the application of those principles cannot be more basic than the principles of justice themselves nor can they come from "outside" the original position.

The Rawls of *TJ* says nothing about the source of principles that guide the application of the principles of justice to the constitution. In *PL*, by contrast,

he is clear about where they come from. The principles of justice and the "guidelines and criteria" of their application are "companion parts of one agreement" (*PL*, p. 226). That agreement is made in the OP (*PL*, p. 225). Even in *PL*, Rawls does not discuss the parties' adoption of the principle of legitimacy and the guidelines of public reason in anything like the detail with which *TJ* discusses their choice of the principles. For example, Rawls gives no indication of the set or menu from which his own guidelines of public reason are chosen, nor does he say what principles of choice are used to decide among the alternatives. It would be good to have the arguments more fully worked out. But what matters for my purposes is that the "guidelines and criteria" are adopted in the original position rather than elsewhere because their purpose is to state the conditions under which power can be exercised to implement the principles of justice. Thus, the guidelines of public reason say what kinds of considerations properly bear on political decision-making. When Rawls asserts an "intimate connect[ion]" between public reason and political legitimacy (*PL*, p. 136), I take it he means that power is exercised legitimately or properly only when its exercise is based on what the guidelines of public reason say are the right kinds of considerations.

Participants in the constitutional convention exercise a form of political power familiar to the contract tradition at least since Locke: constituent power. In a rather enigmatic passage, Rawls says that "the aim of public reason" is to "articulate" the ideal use or expression of that power (*PL*, p. 232). In light of this passage, and of the connection between public reason and the principle of legitimacy, it is not surprising that the liberal principle of legitimacy implies a standard for the proper use of constituent power. It implies that the power to write a constitution is properly exercised only if the essentials of the constitution are such that "all citizens as free and equal may reasonably be expected to endorse them in light of principles and ideals acceptable to their common human reason." For present purposes, I shall take the principle to require that everyone may reasonably be expected to endorse the constitutional essentials in light of the principles and values of justice as fairness. This is just what we would expect if the principle of legitimacy is to guide the application of the principles of justice to the constitution. But what constitutional essentials would free and equal citizens reasonably be expected to endorse in light of the two principles?

As I suggested a moment ago, the guidelines of public reason establish a distinction between reasons that can and cannot justify the exercise of political power—including constituent power—to settle constitutional essentials and matters of basic justice. According to those guidelines, the traditional teachings of Jan's religion do not, as such, provide weighty reasons to violate religious neutrality and cannot justify a constitution which is not neutral. No public reasons, Rawls thinks, could be sufficient to justify a non-neutral constitution. These claims, together with the principle of legitimacy, imply that the only "proper and hence justifiable" (*PL*, p. 217) exercise of constituent power would result in a constitution which is religiously neutral.

I began considering the principle of legitimacy to show how citizens of faith can support a constitution and accept laws that are religiously neutral while affirming their religion's traditional teachings. This question arose because it seemed that Jan could regard a neutral constitution and neutral legislation as justified only if she denied those traditional teachings. It should now be clear that the neutrality of the WOS's constitution does not depend upon the claim that the traditional teachings of a church like Jan's are untrue. It depends instead on the very different claims that those teachings do not, as such, provide weighty reasons to write a constitution that violates religious neutrality, that no public reasons could be sufficient to justify writing a constitution that violates it, and that a constitution arrived at on the basis of nonpublic reasons would require an illegitimate exercise of constituent power. Since Jan's church—and hence Jan—could accept these claims consistent with holding to the truth of traditional teachings, it need not deny those teachings to recognize that a religiously neutral constitution is the only justifiable exercise of constituent power.

The consistency claim can be made vivid if we picture the second stage of the four-stage sequence for choosing and implementing the principles of justice, the viewpoint of the constitutional convention. So far, I have specified that point of view only by reference to the norms that constrain deliberations at the convention, and to the reasons that would be taken into account by the convention's participants. But viewpoints are also specified by the information available in them. The constitutional stage immediately follows the OP in the four-stage sequence. At the constitutional stage, participants have very little information, since the veil of ignorance imposed in the original position is lifted only slightly (*TJ*, p. 197/172). They do not have any information about the truth or falsity of comprehensive doctrines, and so they are in no position to pronounce on the traditional teachings of Jan's religion. That they are not is illustrative. It illustrates the fact that in justice as fairness, conclusions about what exercises of constituent power are justified do not depend upon the denial of such teachings.

The device of the four-stage sequence also helps to show that the same is true of exercises of "ordinary" or legislative power, exercises that result in a religiously neutral body of law. Laws can be thought of as arrived at in the third stage or point of view of the sequence. Since "each point of view inherit[s] the constraints adopted at previous stages" (*TJ*, p. 200/176), the liberal principle of legitimacy constrains the exercise of legislative power.

The argument for this conclusion is straightforward. The constitution lays down procedures for making law and, since the constitution constrains the exercise of legislative power at the third stage, legislation must be enacted according to those procedures. We saw that the application of the principle of legitimacy to constituent power implies that essentials of the constitution must be acceptable to each member of the WOS in light of the principles and values of justice as fairness. It follows immediately that the "exercise of [legislative] 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 light of [those principles and ideals]." This requirement is, of course, just the liberal principle of legitimacy, specified for the exercise of legislative power.

What is meant by saying that legislative power must be exercised "in accordance with" the constitution? The constitution lays down rules for the introduction, passage and interpretation of legislation. It also lays down the limits on legislative power and the ends for which it can be exercised. These ends, and the guidelines for public reason, constrain the exercise of legislative power and interpretive authority, at least when fundamental political issues are at stake. So exercises of such power are "proper and hence justifiable" only when the constraints are honored.

Since I am still concerned with questions raised by citizens of faith, I shall assume that the only exercises of legislative power that satisfy those constraints are those that result in religiously neutral legislation, and that those exercises of power satisfy other conditions on legitimacy as well. If exercises of power are legitimate, the political outcomes that result are legitimate. Rawls sums this up, saying "a legitimate procedure gives rise to legitimate laws and policies made in accordance with it" (*PL*, p. 428). The liberal principle of legitimacy, together with the guidelines of public reason, thus shows how a religiously neutral body of law—like a religiously neutral constitution—can be "proper and hence justifiable" and subject to proper and justifiable enforcement.

The claim that only religious neutral legislation is justified does not depend upon the denial of the traditional teachings of Jan's religion. Like the claim that only a religiously neutral constitution is justified, it depends upon the claims that only some considerations properly bear on the exercise of political power, and that only exercises of such power that are legitimate are justifiable. And so Jan *could* recognize a religiously neutral constitution and religiously neutral legislation as justified without rejecting the traditional teachings of her church. To recognize them as justified, Jan would have to develop what we might call "a sense of legitimacy"—a motivationally effective sense of which considerations provide good reasons for exercising power and of what exercises of power are legitimate. Since Rawls thinks the principle of legitimacy and the guidelines of public reason are chosen along with the principles of justice, it is natural to think of Jan's sense of legitimacy as part of her sense of justice.

This thought seems even more natural if the sense of justice in *PL* is centered on small set of ideal-dependent desires, including the desires to live up to ideals of citizenship and civic friendship and the *Ideal of Democratic Governance*. In §IX.3, we saw that the guidelines of public reason are used to specify the ideal of citizenship and in §IX.4, we saw that they are used to specify the ideal of civic friendship. It should now be clear how those guidelines—together with the principle of legitimacy—are introduced to help to specify the *Ideal of Democratic Governance* as well, and to state conditions

under which the ideal is realized. The connection between this ideal and the guidelines of public reason are confirmed by Rawls's remark that "the aim of public reason is to articulate" "the political ideal of a people governing itself in a certain way" (*PL*, p. 232); the connection between the *Ideal of Democratic Governance* and the liberal principle of legitimacy is confirmed by what seems to be an equation of the *Ideal* with what he calls "the liberal *ideal* of political legitimacy."[4] Once we think of the sense of justice as centered on these ideal-dependent desires, we can see that Jan's sense of justice can move her to acknowledge the justifiability of the laws and constitution of the WOS, and to support its "general structure of authority."

According to this line of thought, citizens of faith in the WOS, such as Jan, can do what many citizens of faith in fact do under the less than ideal conditions of the actual world. They adhere to their church's teachings about a range of matters, including abortion, euthanasia and assisted suicide, the death penalty, and the importance of religious education. But with the experience of living and working and sharing decision-making power with those of different views, they come to believe that the pronouncements of religious authority on these matters do not constitute good reasons for making laws and policies in pluralistic democracy. And they come to believe that it would be improper to enact constitutional provisions or legislation that can only be supported by such pronouncements. Of course, not all citizens of faith believe this in liberal democracies as we know them. But the fact that many do confirms the possibility of a society in which Jan, and most or all others, do.

In sum, Jan's sense of justice would dispose her to support and advocate only exercises of power that the principle of legitimacy says are proper. The basic stability argument then shows how she can affirm this disposition as part of her good from within her comprehensive view. She can do so if, according to her comprehensive doctrine, the values realized when the principle of legitimacy is generally complied with "normally outweigh" the *Goods of Salvific Promotion*. Those values are the values connected with the *Ideal of Democratic Governance*. The principle of legitimacy removes an obstacle that seemed to stand in the way of an overlapping consensus, since it allows Jan's religion to attach sufficient weight to the *Ideal* without denying its own traditional teachings.

But once the obstacle is removed, why would it *actually* attach sufficient weight to them?

In the actual world, citizens of faith who accept the liberal principle of legitimacy may still experience a sense of tension or of loss. They may believe that there are very important goods their society could realize by violating the principle of legitimacy, and significant evils it could avoid by doing so. For

4. Rawls, "Domain of the Political," *Collected Papers*, p. 490.

many, the tension may not ever be finally resolved. And yet their experience of living in a pluralistic society, and abiding by the liberal principle of legitimacy, may make them aware of great social goods that are available only when the principle is widely honored. Their churches may flourish, people of their faith are trusted and recognized as full and equal participants in their societies, and all people are left free to search for God—or not—in their own ways. Their religions may acknowledge, and they may come to believe, that these goods outweigh what is lost. I believe that Rawls counts on the experience of life in the WOS having something like the effect.

Recall that in the congruence arguments of *TJ*, Rawls treated justice as transformative. Joan, the typical citizens of the WOS, who lives under just institutions and develops a sense of justice, becomes the kind of person who values the goods of civic friendship and the expression of her nature above competing goods. That is why she could judge, from within the thin theory, that her balance of reasons tilts in favor of affirming her sense of justice. I believe the Rawls of *PL* thinks that the experience of the WOS, where the principle of legitimacy and the guidelines of public reason are honored, would have a similarly transformative effect on comprehensive doctrines and their adherents. Even if comprehensive views in the WOS do not liberalize in the way that I discussed in the last section, they would attach—or come to attach—the requisite weight to the political values of justice as fairness, including the value of realizing the *Ideal of Democratic Governance*. If this is correct, then Rawls can reach one of the conclusions about Jan that the basic stability argument is supposed to establish, *PL's Nash Claim*:

C_N*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

If Rawls can solve the *mutual assurance problem*, then he can reach:

C_9: "citizens will judge (by their comprehensive view) that political values either outweigh or are normally (though not always) ordered prior to whatever nonpolitical values may conflict with them" (*PL*, p. 392).

From there he can infer:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

Once we see how Rawls can reach these conclusions, we can see "how [it is] possible for citizens of faith to be wholehearted members of a democratic society" and how Rawls would answer what he said is the fundamental question for political liberalism (*PL*, p. xl).

§X.4: Why Political Legitimacy?

According to the reading I developed in the previous section, Rawls relied on the principle of legitimacy to show how justice as fairness can be stable in the face of disagreement about the justice of attempts to implement it in a constitution or in laws and policies.[5] The principle can do this work because exercises of power that might seem to be unjust can still be recognized as legitimate. When they are, they are still seen to enjoy an important form of public justification. A close reading of §31 of *TJ*, on the four-stage sequence, suggests that Rawls tacitly appealed to the notion to help explain how the principles of justice chosen in the first stage are to be applied at the later ones. But the explicit development of this form of justification is new to *PL*.

The importance of legitimacy has not been lost on readers of *PL*. Indeed, one has said that legitimacy is what *PL* is really about.[6] But the role legitimacy plays in the arguments of *PL* and Rawls's reasons for introducing the notion have often been misunderstood. A crucial claim of my reading is that the principle of legitimacy is a principle of right adopted in the original position to guide and judge the *application* of the two principles of justice. This claim is denied by those, such as Charles Larmore, who think the principle of legitimacy is a more fundamental principle that constrains the *adoption* of the two principles as well.[7] In the conclusion, I shall look closely at Larmore's interpretation, for doing so brings to light an important point about the justification of political liberalism.

In §VIII.6, we saw that Burton Dreben thought Rawls made the turn to political liberalism because he came to recognize that under conditions of pluralism, citizens of a WOS might disagree with the principles of justice. Disagreement about the standards of justice, in turn, might seem to be the reason citizens would differ in their judgments of constitutions and outcomes, and so might seem to be the reason Rawls needed the principle of legitimacy. This is Dreben's view. As I argued that Dreben is mistaken about Rawls's reasons for taking the political turn,[8] so I also believe he is mistaken about Rawls's reasons for introducing the principle of legitimacy. For even citizens who accept the same principles of justice, and who agree about what political values bear on constitutional essentials and matters of basic justice, might still disagree about how those values are to be balanced in particular cases. In *TJ*'s discussion of the four-stage sequence, Rawls implies that this will be true in

5. The arguments of this section are expanded and clarified in my "Legitimacy and the Project of *Political Liberalism*" (forthcoming).

6. David Estlund, "The Survival of Egalitarian Justice in John Rawls's *Political Liberalism*," *Journal of Political Philosophy* 4 (1996): pp. 68–78, p. 68.

7. Larmore, *Autonomy of Morality*, pp. 147ff.

8. See Chapter VIII, note 32 and the accompanying text on Dreben's interpretation.

many cases (*TJ*, pp. 199–99/174). In *PL*, he explicitly says that this is true in the case of abortion (see *PL*, p. 243, note 32). When it is true, citizens will disagree about the justice of legislation. If those who do not prevail are to support the institutions under which they live even when they believe those institutions have made a terrible mistake, they must recognize that legislation with which they disagree is nonetheless legislation that was properly enacted. The principle of legitimacy is therefore needed to show—contra Dreben—how stability is possible even when citizens agree on principles of justice.

Why didn't Rawls introduce the principle of legitimacy in *TJ*? Why did he explicitly appeal to it only in *PL*? In §§III.3 and VIII.5, we saw that Rawls thinks *TJ*'s treatment of stability covers the "simplest case."[9] It did not occur to him then that justice as fairness might need to be stabilized by an overlapping consensus. Nor did he look very deeply into why citizens might disagree about the justice of outcomes reached at the constitutional and legislative stages. By the time he wrote *PL*, Rawls recognized that if justice as fairness is to be stable, citizens must regard it as justified from within their comprehensive doctrines. He also realized that some citizens might dispute the justifiability of attempts to implement justice as fairness at the constitutional and legislative stages, because they thought the outcomes of those stages were at odds with their comprehensive views. Rawls relied on a new and powerful form of justification—legitimacy—to show how citizens could regard the outcomes as justified without denying the relevant parts of their comprehensive doctrines. Thus, it was a deeper appreciation of the role of comprehensive doctrine in the stability and justification of justice as fairness, not the possibility that some citizens of a WOS would reject justice as fairness, that moved Rawls to make one of the important changes between *TJ* and *PL*: the introduction of political legitimacy.

Is belief in the legitimacy of laws and institutions enough to secure wholehearted allegiance to a liberal society? The word "wholehearted" can suggest that all the citizens of the WOS have a great deal of unmitigated positive affect for their basic institutions. But showing C_9 and C_{PL} does not demonstrate the presence of such affect. These conclusions show that citizens judge that it is right and, on balance, good not to defect from the basic terms of cooperation as those are institutionalized in the "general structure of authority" (*PL*, p. 136). They can reach these judgments while being so deeply distressed by some of the laws their society enacts that they find an admixture of bitterness in their affect for their society.

Consider Rawls's two examples: Quakers who "refuse to engage in war" and those who oppose the legality of abortion (see *PL*, pp. 393–94, and 394, note 32). These citizens adhere to comprehensive doctrines according to which war and abortion are unjust. In these cases, as in the example of Jan, whose religion teaches that it offers the surest way to salvation, Rawls appeals to the

9. Rawls, "Political not Metaphysical," *Collected Papers*, p. 414, note 33.

principle of legitimacy to show that they can accept the "general structure of authority" without denying the teachings of their religions. For if these citizens accept the liberal principle of legitimacy, they can recognize the decision to go to war or to legalize abortion as "legitimate (even if not just)" (*PL*, p. 394). And if their faiths teach that a constitutional regime in which the principle is generally honored realizes important goods, "allegiance to a just constitutional government may win out within the religious doctrine[s]" (*PL*, p. 394). Then (9.2) and (9.5) will be true of the doctrines, and *PL's Nash Claim*, C_9, and C_{PL} will be true of its adherents. They will be true even if these citizens still think their society is very badly tainted by its decisions to take, or to allow the taking of, human life or innocent human life.

The actual world is not, of course, the ideal world of justice as fairness. The account of stability developed for the latter may not explain—nor is it intended to explain—such stability as is enjoyed in the former. In the actual world, some citizens of faith are deeply distressed by the fact that the liberal democracies in which they live often reach what they regard as the wrong decisions on issues, such as abortion, that they regard as among the most important they face. Their distress may not lead them to give up on constitutional government. But instead of developing wide-ranging ties of civic friendship with others, they may increasingly regard themselves as "resident aliens" in their society. They may have or develop serious doubts about the liberal principle of legitimacy and the *Ideal of Democratic Governance*, and be quite ready to give their support to a form of constitutional democracy that is not based on a political liberalism.[10] If this is so, then the stability enjoyed by actual liberal democracies may be quite different from the stability that would prevail in ideal theory. It may, indeed, be closer to a *modus vivendi* than to the overlapping consensus that would obtain in the WOS of justice as fairness.

§X.5: A Question about the Arguments for C_9 and C_{PL}

In Chapter II, I argued that Rawls was concerned to show justice as fairness would not be destabilized by "the hazards of the generalized prisoner's dilemma" (*TJ*, p. 577/505). To show that it would not be, and that it would be inherently stable, Rawls saw that he needed to establish a Nash claim. He needed to show, roughly, that planning to maintain a sense of justice is each person's best reply to the similar plans of others (*TJ*, p. 468/497). In Chapter VIII, we saw that Rawls's arguments for *TJ's Nash Claim* C_N failed because of the

10. I discuss this possibility more fully in my review of Jeffrey Stout, *Democracy and Tradition* (Princeton, NJ: Princeton University Press, 2004); the review can be found at *Faith and Philosophy* 23 (2006): pp. 221–29. The phrase "resident aliens" is Stanley Hauerwas's; see Stanley Hauerwas and William Willamon, *Resident Aliens: Life in the Christian Colony* (Nashville, TN: Abingdon Press, 1989).

pluralism of comprehensive doctrines that would be found in a WOS. We have now seen how the Rawls of *PL* reconstructed justice as fairness so that he could move from (9.1), the claim that members of the WOS follow their comprehensive doctrines, to *PL's Nash Claim* C_N^*.

As in *TJ*, so in *PL*, establishing a Nash claim is not enough to show stability. Stability requires that everyone actually regulate her plans by her sense of justice. And so the Rawls of *PL* wants to move from C_N^* to:

C_9: "citizens will judge (by their comprehensive view) that political values either outweigh or are normally (though not always) ordered prior to whatever nonpolitical values may conflict with them" (*PL*, p. 392).

And he wants to move from C_9 to:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

At the end of §X.3, I said that to reach the conclusions he wants, Rawls needs to solve the *mutual assurance problem.*

Even if the explanations I have offered for various features of political liberalism so far have been convincing, it may seem to this point that an argument for *PL's Nash Claim*—and hence a concern with a problem of mutual assurance–can only be located in Rawls's texts only by forcing *PL* into a pattern of argument I found in *TJ*. It is true that Rawls's later essays offer little explicit evidence that he was concerned with the *mutual assurance problem*. There is, however, one telling indication of his concern in "The Domain of the Political and Overlapping Consensus." There Rawls says that showing stability requires showing that citizens of the WOS would develop "a sufficiently strong sense of justice guided by appropriate principles and ideals, so that they normally act as justice requires, *provided they are assured that others will act likewise*."[11] It is, I think, a merit of the reading of *PL* offered here that it shows exactly where the *mutual assurance problem* referred to in this passage arises.

I suggested in §IX.1 that Rawls solves that problem by appealing to the existence and public knowledge of an overlapping consensus. For when an overlapping consensus obtains—and we saw in §X.3 why Rawls thinks it would in a WOS—no one's conception of the good provides sufficient reason to act against the demands of justice. And when an overlapping consensus is also known to obtain, each person has the assurance that no one else's conception of the good provides sufficient reason to act against those demands. As I showed in §IX.1, the claim that this is in fact how Rawls solves the *mutual assurance problem* is a claim that enjoys some textual support. It gets that

11. Rawls, "Domain of the Political," *Collected Papers*, p. 479 (emphasis added).

support from the facts that Rawls can move to C_9 only by solving the *mutual assurance problem* and that in *PL*, he explicitly infers C_9 from the existence and public knowledge of such a consensus (*PL*, p. 392).

In §§VI.5 and VII.3, we saw that the Rawls of *TJ* treats mutual assurance as if it went without saying. As Rawls came to appreciate the pluralism that forced his political turn, he also came to appreciate that that problem was considerably more complicated than he had previously assumed. For he came to see that pluralism makes it difficult to establish one of the facts that is supposed to solve the *mutual assurance problem* in *PL*: the fact that in a WOS, an overlapping consensus not only obtains but would be known to obtain. Solving the *mutual assurance problem* required the Rawls of *PL* to deploy conceptual resources that had played a minor role in *TJ*. To see the complications, let us consider an example.

The basic stability argument presupposes that members of the WOS have a sense of justice. They presuppose, that is, that Rawls has established:

C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and society included in justice as fairness.

So far, I have taken C_3' to mean that all the members of the WOS want to live up to those political ideals *as such* or *under those descriptions*. So, for example, I have taken it to mean that all members of the WOS think of themselves as free and equal citizens and think—as (1.9') says—that the conceptions of freedom and equality which best specify their citizenship are the conceptions represented in the OP. I have also taken it to imply that citizens of the WOS all want to live up to that ideal of themselves as persons who act from the principles that would be chosen there. I do not mean that they all consciously entertain these ideals and desires. I mean, rather, that they can properly be ascribed to them as the best explanations of their deliberations and their conduct toward others, including the justifications they offer others for their conduct. And I have supposed this to be true of citizens who affirm comprehensive doctrines that they recognize to have political and social implications, such as Kantianism, utilitarianism, and various religious views. I now want to question this last supposition, since I suspect that it misdescribes the sense of justice possessed by adherents of some fully comprehensive doctrines.

Let us call a comprehensive doctrine *very fully comprehensive* if it meets the following conditions. It includes norms, values, and ideals for all subjects. It has versions of those concepts that are worked out specifically for political institutions, and so it covers the domain of the political, but it does so using concepts and values many of which are quite different from those of justice as fairness. And it is borne by associations that are capable of imparting the view very effectively. I cannot spell out such a view in any detail, but here is a brief sketch.

Imagine a religious view according to which various social forms, including political life, aim at the common good of participants in those forms. Because human beings are thought to have a powerful need for belonging to social

entities larger than themselves, in which all flourish, the common good includes the flourishing of relations among members of society. The most flourishing relation is that of mutual love, which can be differently realized in different social forms. Rights and liberties are understood as the minimal conditions of life in political community, because life in a political community is impossible if some are permitted to dominate others. Some rights and liberties are also thought to be necessary to so that all members of political society can search for God freely, so the religious doctrine is one of free faith. The doctrine endorses provision of a social minimum because living in poverty is beneath the dignity of persons made in God's image. So the religious view under consideration holds out an ideal of political society as an egalitarian community of tolerance and mutual love, and an ideal of the individual as a child of God made for participation in that—and other—communities.

In imagining a society that includes this *very fully comprehensive view*, I am not imagining one in which the educational forces that Rawls says are at work in a WOS—and that bend comprehensive views toward justice as fairness—fail completely. For it may be that the public conception has bent the *very fully comprehensive doctrine* to itself in important ways. Over time, that doctrine has come to recognize the importance of toleration and other human rights and to advocate for them forcefully, and has come to acknowledge the importance of egalitarian economic and political arrangements. It has ceased to view liberal democracy as a threat, and now defends it in official pronouncements. It does not claim that political power should be used to encourage adherence to comprehensive doctrine, and its adherents think political questions should be settled by values associated with the political common good. But the forces of social learning have not resulted in the doctrine's incorporating the ideals of justice as fairness as such, since the doctrine is worked out and taught in its own terms.

It is not at all clear how the view I am imagining could be developed beyond the sketch I have presented. The notion of the common good, in particular, is vague and in need of considerable development. But pluralistic societies contain many comprehensive doctrines that attract wide followings despite their lack of rigorous philosophical underpinnings. What matters for present purposes is that if the view I have imagined were present and effectively fostered in a society the public conception of which was justice as fairness, then some members of that society would have conceptions or ideals of themselves and their society that are very different from the conceptions included in the public conception. They might well have a powerful sense of justice, but that sense of justice would—I am imagining—be informed by different ideals, values, and concepts than those referred to by C_3'

I assume, then, that the ideal-dependent desires to which C_3' refers cannot be ascribed to adherents of the comprehensive doctrine I have imagined, so that C_3' is false. But I am also assuming that all citizens have senses of justice that lead them to support liberal and egalitarian outcomes. Indeed, let's suppose that adherents of the *very fully comprehensive view* all support the same

political outcomes—or almost all the same political outcomes—as are supported by citizens whose sense of justice is informed by justice as fairness. Then the society to which they all belong will be just and, I assume, stably just. If we are reluctant to describe that society as a WOS because C_3' is false of it, then it remains true that it is a reasonably just liberal democracy that would be stable. But is its stability accounted for by an overlapping consensus on justice as fairness, as I have so far taken the Rawls of *PL* to suppose?

If I am right that an overlapping consensus obtains in a society just in case (9.2) and (9.3) true of it, then the answer depends upon whether the *very fully comprehensive view* is "either congruent with, or supportive of, or else not in conflict with" the values and ideals of justice as fairness. If I was right about what cases Rawls has in mind when he speaks of comprehensive doctrines being "congruent with" justice as fairness, then it seems clear that the *very fully comprehensive view* is not such a case. Moreover, since adherents of the view endorse the same political outcomes as those whose sense of justice is informed by justice as fairness, the view may be "supportive of, or else not in conflict with" justice as fairness *in some way*. But given the great conceptual differences between the two, it is hard to see how the view is "supportive of, or else not in conflict with" *the values and ideals of justice as fairness*. So it would seem that the answer to my question is "no." In that case, (9.3) is false of the society I am imagining, and Rawls cannot infer

(9.6) Each member of the WOS judges, from within her comprehensive view, that the political ideals referred to by C_3' and the values of realizing them "normally outweigh whatever values are likely to conflict with them," at least when others reach the same judgment.

Without (9.6), it is hard to see how Rawls could move from (9.2) and (9.3) to *PL's Nash Claim* C_N^*. And so it is hard to see how he could move from (9.2) and (9.3) to C_{PL}. And since it is hard to see how he could do that, it seems that his argument that an overlapping consensus on justice as fairness is what stabilizes a just liberal society fails.

But while Rawls says in (9.3) that reasonable comprehensive doctrines are "either congruent with, or supportive of, or else not in conflict with" the values and ideals of justice as fairness, he is not very specific about the "supportive of" relation. The lack of specificity is a matter of principle, since—as we saw at the end of Chapter IX—the relation between comprehensive doctrine and political conception is left to citizens to work out for themselves (*PL*, p. 11). Perhaps adherents of the *very fully comprehensive view* can see their view as fitting with justice as fairness in such a way that (9.3) and (9.6) are true. Then Rawls could get as far as *PL's Nash Claim* after all:

C_N^*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

The mere *existence* of an overlapping consensus, however, does not get Rawls far enough. Even if he can get to C_N*, he still needs to solve the *mutual assurance problem* in order to reach C_9 and C_{PL} and show stability. The phrase "at least" in C_N* leaves open the possibility that some citizens of the WOS will affirm their sense of justice even if they do not know whether others affirm theirs. I have included the phrase because, as we shall see, Rawls explicitly allows for this possibility. But even these citizens may act from the principles of right chosen for non-ideal conditions if they think their fellow citizens are not committed to justice. Even they may want to be assured that everyone else is just before they try to live up to the ideals of justice as fairness by acting from the principles of ideal theory instead.

The WOS is and is known to be just. In *PL* as in *TJ*, what members of the WOS want to be assured of is that everyone else's commitment to justice as fairness really is "wholehearted." For if each has this assurance, each knows that no one else's view of what is good would move him to defect from justice as fairness when he thinks he can evade punishment or when the balance of power changes to his advantage. But adherents of the *very fully comprehensive doctrine* that I have imagined reason and speak about justice quite differently than their fellow citizens do. While they support the political outcomes justice as fairness supports, they explain their support for those outcomes by talking about a common good, mutual love, conditions of life in community, and the right to search for God freely. How, then, are their fellow citizens to know that their commitment to *justice as fairness* is wholehearted?

(9.2) says that when an overlapping consensus obtains, comprehensive doctrines endorse justice as fairness; when an overlapping consensus is known to obtain, those doctrines are presumably known to endorse it. If it were common knowledge that the *very fully comprehensive doctrine* was part of an overlapping consensus, everyone would know that the doctrine endorses justice as fairness. According to (9.1), citizens follow their comprehensive doctrine and the fact that they do would also presumably be common knowledge in a WOS. So if it were common knowledge that the *very fully comprehensive doctrine* was part of an overlapping consensus, it would also be common knowledge that adherents of the doctrine do not have reason to defect from justice as fairness and, indeed, find reasons to endorse their sense of justice in their comprehensive view. This piece of common knowledge would provide others assurance of their wholeheartedness. In that case, public knowledge of the existence of an overlapping consensus would solve the *mutual assurance problem*, just as Rawls implies. But when the concepts used by the *very fully comprehensive doctrine* are so different from those of justice as fairness, how can it be known to take part in an overlapping consensus? How is public knowledge of such a consensus possible?

This difficulty cannot be avoided by relaxing the assumption expressed by (9.2) and (9.3), according to which a WOS is characterized by an overlapping consensus on justice as fairness. Rawls does eventually relax that assumption, and grant that there is more likely to be an overlapping consensus on "a class of liberal conceptions that vary within a more or less narrow range" (*PL*,

p. 164). As I shall explain in the next section, I believe his concession was motivated in part by the possibility of *very fully comprehensive doctrine*. But this concession does not mitigate the difficulty of the *mutual assurance problem.* For even if the *very fully comprehensive doctrine* is or implies one of the "liberal conceptions," and so takes part in an overlapping consensus, those who do not adhere to the doctrine still need assurance that it does. They need it because they still need to know that adherents of the *very fully comprehensive doctrine* are wholeheartedly committed to a reasonable liberal political conception that supports, for example, basic rights and liberties. Thus even if (9.2) and (9.3) are weakened, so that they assert an overlapping consensus on a class of liberal conceptions, the facts that there is such a consensus, and that it includes the *very fully comprehensive doctrine*, still need to be publicly known. How they can be publicly known when adherents of the *very fully comprehensive doctrine* reason about justice using concepts of their own remains to be seen.

§X.6: Public Reason, Mutual Assurance, and Pluralism about Justice

One of the notable differences between *TJ* and *PL* is the prominence of public reason. In spelling out his account of public reason, Rawls defends guidelines for debate of fundamental questions in the public forum. In §§IX.2 and IX.3, when I discussed political ideals of conduct and society, we saw some reasons for the importance of those guidelines. The prominence of public reason in *PL* is also explained in part, I believe, by the need to solve the *mutual assurance problem*. I suggest that Rawls developed the guidelines of public reason with that problem in mind.

Rawls's treatment of public reason has generated an enormous body of literature; I cannot engage it here. Instead, I shall say just enough about the development of Rawls's views of public reason to lend some credence to my suggestion. I shall begin by looking more deeply into (9.2), the claim that "reasonable doctrines endorse the political conception, each from its own point of view" (*PL*, p. 134), to see what reasonable comprehensive doctrines endorse the political conception for. For simplicity's sake I continue to assume, for now, that a WOS is characterized by an overlapping consensus on a single conception of justice, justice as fairness.

In a WOS, the public conception of justice provides a "common point of view" (*TJ*, p. 5/4) or a "unified perspective" (*TJ*, p. 474/415) in which the settlements of citizens' competing claims are "adjudicated" (*TJ*, pp. 5, 474/4, 415). In §II.3, we saw that points of view are defined by rules of reasoning and information drawn on by those who occupy them. When Rawls says that the public conception of justice provides a point of view for "adjudicat[ing]" citizens' competing claims, I take it he means the conception furnishes values and principles on the basis of which questions of basic justice are to be settled, and rules of reasoning for moving from those values and principles to a settlement.

For citizens of the WOS to "acknowledge" their "common point of view" (*TJ*, p. 5/4) is for them to acknowledge that political outcomes are justifiable only if they can be supported by the values and principles of the political conception. When political outcomes meet that standard of justifiability, they satisfy the liberal principle of legitimacy and the WOS realizes the *Ideal of Democratic Governance.* When an overlapping consensus obtains and comprehensive doctrines "endorse the political conception," they endorse it as providing the values and principles by which political outcomes are to be justified. If an overlapping consensus is also known to obtain—so that everyone knows that all reasonable comprehensive doctrines acknowledge that the public conception provides the requisite point of view—then this fact about comprehensive doctrines will itself be common knowledge. I assume that citizens not only follow their comprehensive doctrines, as (9.1) says, but are commonly thought to do so. So when an overlapping consensus is known to obtain, it is also common knowledge that citizens accept the authority of that point of view on the basis of their comprehensive doctrine. Their acknowledgement of that point of view as authoritative is then known to be wholehearted.

Appeal to the existence and public knowledge of an overlapping consensus is therefore sufficient to solve the *mutual assurance problem*. Whether it is necessary depends, in part, upon what concepts and methods of reasoning citizens of the WOS actually use when they argue about basic political questions. Rawls says that he was initially drawn to what he calls the "exclusive view of public reason" (*PL*, p. 247, note 36). This is the view that citizens should never introduce reasons drawn from comprehensive doctrines into public debate about fundamental questions (*PL*, p. 247). According to the exclusive view, the only reasons that may be brought to bear are those provided by the values and ideals of the political conception of justice. To comply with the exclusive view just is to reason about questions exclusively from the "unified perspective" provided by that conception.

The exclusive view is highly restrictive. Part of its attraction, I believe, was that it promised an elegant solution to the assurance problems that can be posed by *very fully comprehensive doctrines*. If citizens use the concepts of their *very fully comprehensive doctrine* to debate basic political questions, their arguments may suggest that they do not acknowledge the authority of the political conception to adjudicate them. But if all the members of the WOS—including adherents of *very fully comprehensive doctrines*—comply with the exclusive view, then they all adopt and are known to adopt the "common point of view" or "unified perspective" whenever basic political questions are at issue. So long as they can be assumed sincere, the way they reason about these questions in public confirms their allegiance to justice as fairness and the *mutual assurance problem* does not arise. The solution promised by the exclusive view depends upon the existence of an overlapping consensus, since citizens might not comply with the requirements of the view unless their comprehensive doctrines endorsed justice as fairness. But given the existence of an overlapping

consensus, it seems to solve the *mutual assurance problem* directly, without appealing to public knowledge of a consensus.

Despite the attraction Rawls felt for the exclusive view, he never endorsed it. One of the reasons he did not, I think, is that he recognized that the view could not make good on its promise to eliminate the *mutual assurance problem* directly.

Divisions about some political questions—Rawls's example is the question of whether church schools should receive public funding (*PL*, p. 248)—can be so deep that adherents of different comprehensive doctrines come to doubt one another's allegiance to political values. Rawls does not spell out the example in any detail. I presume what he has in mind is that even if champions of public funding publicly defend their position by appealing only to the political values of religious equality and religious liberty, their argument raises questions about whether they are also committed to church–state separation. Perhaps, it will be thought, they are using political values as a cover and do not really acknowledge the authority of those values. So the *mutual assurance problem* can arise even when citizens of the WOS comply with the exclusive view. "One way this doubt may be put to rest," Rawls suggests "is for the leaders of the opposing groups to present in the public forum how their comprehensive doctrines do indeed affirm [the] values [of the public conception]" (*PL*, p. 249). This is, in effect, the suggestion that leaders of opposing groups make the existence of an overlapping consensus publicly known. Once the existence of an overlapping consensus is publicly known, Rawls thinks, the sincerity of each side's appeals to political values will no longer be in doubt. Mutual assurance of sincere and wholehearted allegiance to the political conception is therefore provided—as Rawls implied it would be when he argued for C_9 in "Reply to Habermas"—by appeal to public knowledge of an overlapping consensus.

But Rawls quickly came to think that even the inclusive view was too restrictive. By his last published treatment of public reason, in "Idea of Public Reason Revisited," Rawls famously endorsed what he called the "wide view." The wide view allows citizens to introduce their comprehensive doctrines into public political argument at any time, subject to one restriction I shall mention below. Some readers have thought that in moving from the exclusive to the wide view, Rawls moved from a view of public reason that was overly confining to one that is too permissive. Charles Larmore, for example, writes that "in the forum where citizens officially decide the basic principles of their political association and where the canons of public reason therefore apply, appeals to comprehensive doctrine cannot but be out of place . . . at least in a well-ordered society."[12] But the wide view can, I believe, be defended if we see the account of public reason is framed to solve the *mutual assurance problem*, since the wide view—though permissive—is strong enough for that.

12. Charles Larmore, "Public Reason," in *The Cambridge Companion to Rawls*, pp. 368–93, p. 386f.

The wide view applies only to ordinary citizens. Whether the wide view is too permissive depends upon why ordinary citizens are subject to guidelines of public reason at all. Larmore suggests that there are times when ordinary citizens "officially decide the basic principles of their association," and that the guidelines of public reason apply to them because of the decisions they have to make. His idea seems to be that, as we might think comprehensive views are out of place when government officials decide basic questions, so they would be out of place when ordinary citizens decide them too. Rawls agrees with Larmore about public officials, for he says that judges, elected officials, and candidates for public office are subject to more stringent restrictions on public reason.[13] Ordinary citizens vote for public officials and one of the ways they live up to the idea of public reason is by holding officials responsible for conforming to the guidelines that apply to *them*.[14] But Rawls implies that ordinary citizens rarely if ever "officially decide the basic principles of their political association" themselves.[15] There must be another reason—different than the one cited by Larmore—that ordinary citizens are subject to "the canons of public reason."

As my remarks will already have suggested, I believe the answer is this: Members of the WOS all need assurance that everyone else acknowledges the authority of the "unified perspective" on fundamental questions that the political conception provides. Rawls's concern with this assurance problem, rather than with the question of how citizens themselves settle basic political questions, would explain the content of the wide view. For the wide view allows citizens to introduce comprehensive doctrines into public political discussion—and, presumably, to vote—on the basis of their comprehensive doctrine "provided that in due course public reasons, given by a reasonable political conception, are presented sufficient to support whatever the comprehensive doctrines are introduced to support" (*PL*, pp. xli–xlii). When I discussed the exclusive view, I said that to reason about political questions using exclusively public reasons is to adopt and reason from citizens' "common point of view." So the wide view allows citizens to introduce and base their votes on comprehensive doctrine, provided that in due course they adopt and reason from that common viewpoint as well.

Rawls refers to the "provided that" clause as "the proviso." The difficulty with interpreting it is figuring out what Rawls means by "in due course." On my reading, Rawls allows citizens to rely on their comprehensive doctrines—including *very fully comprehensive doctrines*—without adducing public reasons in support of their positions, so long as their doing so does not lead others to doubt that they acknowledge the authority of the public conception of justice. If doubts never arise, then the proviso is never triggered and they

13. Rawls, "Public Reason Revisited," *Collected Papers*, p. 575.
14. Rawls, "Public Reason Revisited," *Collected Papers*, p. 577.
15. Rawls, "Public Reason Revisited," *Collected Papers*, p. 577.

need do nothing more. Only if doubts arise, and others need assurance of their allegiance, must they provide assurance by actually adopting and reasoning from the "unified perspective" the public conception of justice provides. That is, I believe, why Rawls says that "the details about how to satisfy [the] proviso must be worked out in practice and cannot feasibly be governed by a clear family of rules given in advance".[16]

The restrictions of public reason are sometimes said to show that Rawls is deeply suspicious of comprehensive doctrines, especially religious ones, or that he thinks religious political argument is inherently destabilizing. If my interpretation is correct, these are serious misreadings, at least of the wide view that Rawls endorsed in his last writings on public reason. The wide view allows reliance on religious political argument at any time, restricted *only* by the proviso. As I have just tried to show, the motivation for the proviso is not the conviction that religion destabilizes society or that it leads to civil strife. It is the fact that a person's reliance on religious argument can lead her interlocutors to doubt whether she acknowledges the political authority of justice as fairness. Rawls could have required citizens to assure one another of their commitments by requiring them to comply with more restrictive guidelines of public reason than those associated with the wide view. He could, for instance, have argued that citizens must *preempt* others' doubts about their acceptance of the political conception. In that case, he might have replaced the phrase "in due course" in the proviso with the phrase "at the same time." Instead, the proviso requires citizens to adopt and deliberate in their "common point of view" only when they have good reason to think assurance is actually needed. In defending it, Rawls advocates what is, by construction, the weakest and least restrictive guideline sufficient to solve the *mutual assurance problem*.

We have seen why Rawls thinks that problem needs to be solved. Showing that justice as fairness would be inherently stable—or, as Rawls says in his later work "stable for the right reasons"–is critical, for reasons we saw at the end of Chapter II. It will enjoy the right kind of stability only if everyone in the WOS knows that everyone else is committed to living up to its values and ideals. To claim that even the wide view is too restrictive is, in effect, to favor stability of some other sort than the kind Rawls wants to show.

Members of the WOS will be motivated to satisfy the proviso, and to provide others assurance of their commitment to the political conception, only if they really are wholeheartedly committed to it. They will be committed to it only if their comprehensive doctrine supports it. More precisely, they will be committed to it only if it is true that:

(9.2) "Reasonable doctrines endorse the political conception, each from its own point of view" (*PL*, p. 134).

16. Rawls, "Public Reason Revisited," *Collected Papers*, p. 592.

(9.2) expresses Rawls's assumption that an overlapping consensus obtains, an assumption vindicated by the arguments reviewed in §X.2. So the effectiveness of the wide view at solving the *mutual assurance problem* presupposes the existence of such a consensus. I have said that Rawls solves that problem by appealing to both the existence and public knowledge of an overlapping consensus. How does the success of the wide view depend on this public knowledge?

I believe Rawls wants to leave open the possibility that citizens of the WOS can often use their comprehensive doctrines to argue about political questions without having to assure their interlocutors of their commitment to justice as fairness by adopting citizens' "common point of view." That will be possible only if they can rely on their comprehensive doctrine without raising doubts about their commitment. Whether *that* is possible is quite likely to depend upon the knowledge of one another's comprehensive doctrines that citizens bring to politics. More specifically, it is quite likely to depend upon everyone's knowing that everyone else's comprehensive doctrines endorse justice as fairness and that an overlapping consensus obtains.

Common knowledge of an overlapping consensus is not based only, or even primarily, on what citizens say issue-by-issue in the public forum. Rather, it is part of the knowledge of a society's political culture that citizens build up over time. While in some societies, citizens may know little about one another's doctrines and reliance on them may easily arouse suspicion, in others citizens may learn enough to assure themselves of one another's commitment to the values and ideals of the political conception. That is why Rawls says that how the proviso is to be satisfied "is determined by the nature of public political culture and calls for good sense and understanding".[17] Thus it is also by transmission of the public political culture that knowledge of an overlapping consensus becomes public.

So far, I have supposed for simplicity's sake that a WOS is characterized by an overlapping consensus on a single conception of justice, justice as fairness. To see whether this simplifying assumption is realistic, let us return to the example of the *very fully comprehensive doctrine* introduced above. I said that adherents of this doctrine support many of the political outcomes that adherents of justice as fairness support, and think that those outcomes should be reached on the basis of political values, but that they have a sense of justice which is informed by different values and ideals than those of justice as fairness. They reason about politics using different concepts, except when fulfilling the proviso requires them to do otherwise.

Now suppose that a society is otherwise well-ordered by justice as fairness, but includes this doctrine. It may be that members of the society who do not adhere to it will believe, on the basis of background and contextual knowledge and the willingness of adherents to satisfy the proviso, that the *very fully comprehensive doctrine* supports justice as fairness. And so it may be that they all believe an overlapping consensus on justice as fairness obtains. But a more natural way

17. Rawls, "Public Reason Revisited," *Collected Papers*, p. 592.

to describe the case, I think, is to say that they believe adherents of the *very fully comprehensive doctrine* all endorse a liberal political conception of justice that supports the outcomes justice as fairness supports. If this more natural description is the right one, then members of this society do not believe (9.2), which says that "reasonable doctrines endorse *the* political conception, each from its own point of view" (*PL*, p. 134, emphasis added). And so they do not all believe that there is an overlapping consensus *on justice as fairness* after all.

I have assumed that an overlapping consensus is adequately described by (9.2) and by the state of affairs described in (9.3):

(9.3) Each comprehensive doctrine is "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 for a democratic regime" (*PL*, p. 169).

But these are not Rawls's only descriptions of an overlapping consensus. The claim that there would be widespread belief in an overlapping consensus on justice as fairness when some people accept a *very fully comprehensive doctrine* seems even less plausible when we turn to another description.

In the revised *Dewey Lectures*, Rawls says that when an overlapping consensus obtains, "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*, p. 38). Perhaps Rawls could insist that the conception of justice associated with the *very fully comprehensive doctrine* "can be seen . . . to coincide with" justice as fairness. But this will be plausible only if "coincide[nce]" of conceptions of justice is equated with coincidence on outcomes, and that seems strained interpretation of "coincide." It seems less strained and more accurate to grant that citizens do not believe they all endorse coincident conceptions of justice. What they believe is that they endorse different conceptions of justice that coincide on basic political questions and on their support for just institutions. Because of these coincidences, I assume that the society I am imagining will be stably just. The possibility of a stably just liberal society in which some citizens accept a *very fully comprehensive doctrine* raises the question of whether its stability needs to be accounted for by an overlapping consensus on a *single* conception of justice. If the society I am now imagining is one in which there is not widespread belief in the existence of such a consensus, then the answer is "no."

How diversity about justice would come about in a WOS is too complicated a matter fully to explore here. But it may be that the inability of Rawls's account to handle cases like the one I have just imagined is *one* of the reasons he changed his characterization of a WOS. In *TJ*, a WOS is said to be a society that is "effectively regulated by *a* conception of justice" (*TJ*, pp. 5/4–5, emphasis added). In *PL* he says that such a society is possible, but that consensus is more likely to be on a "class of liberal conceptions that vary within a more or less

narrow range" (*PL*, p. 164) than on a single conception. The society in which everyone accepts justice as fairness, once front and center in discussions of stability, is now treated as an example (*PL*, p. 164) or a limit case (*PL*, p. 167). Rawls does talk about a society well-ordered by justice as fairness as a limit it would be desirable to reach. This, I suspect, is because, even while acknowledging that reasonable people can differ about which liberal political conception they endorse, Rawls cannot help showing that he thinks justice as fairness is the most reasonable (*PL*, p. xl). But if an overlapping consensus on justice as fairness is unrealistic or unlikely, then Rawls would have to concede that even

C_3': All members of a WOS want to live up to the political ideals of conduct, friendship, and association included in justice as fairness.

is too strong. Much more plausible is:

C_3'': All members of a WOS want to live up to the political ideals of conduct, friendship and association included in any of a family of liberal political conceptions of justice.

If this is correct, then some of the steps in *PL*'s basic stability argument will have to be slightly revised. (9.2), which says that "reasonable doctrines endorse *the* political conception," will have to be amended to read:

(9.2') "Reasonable doctrines endorse [a liberal] political conception, each from its own point of view" (*PL*, p. 134).

Steps (9.4), (9.5), and (9.6), which refer to C_3', will have to be recast so that they refer to C_3'' instead. *PL's Nash Claim*, will also have to be recast. It says:

C_N^*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

The new version will have to replace the phrase "the values and ideals of justice as fairness" with "the values and ideals of a reasonable political conception of justice," and the phrase "those values and ideals" with "such values and ideals." Finally, I have so far taken the conclusion of the basic stability argument to be:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

That conclusion will have to be revised so that it does not refer specifically to justice as fairness.

We have now seen that the pluralism of the WOS undermined Rawls's argument for *TJ's Nash Claim* and forced him to defend *PL's Nash Claim* instead. We have also seen that even if Rawls can get to the latter claim, he still needs to solve the *mutual assurance problem* if he is to move from that claim to

C_9 and C_{PL}, and show that justice as fairness would be stable in the right way. I have suggested that pluralism—in the form of *very fully comprehensive doctrines*—also complicates the *mutual assurance problem*. I suggested at the beginning of this section that we read Rawls's account of public reason as motivated, in part, by a desire to solve that problem in its more complicated form. We have seen that my reading explains, not only Rawls's initial attraction to the exclusive view of public reason, but also his rejection of it in favor of a more flexible and inclusive view. We have also seen that it explains why Rawls came to think that the relatively weak wide view of public reason is still strong enough to "secure[] what is needed" (*PL*, p. lii). My interpretation draws additional textual support from the fact that the weak view's success at solving the *mutual assurance problem* depends upon both the existence and the public knowledge of an overlapping consensus—just as we would expect, given the grounds on which Rawls infers C_9 in his "Reply to Habermas."[18]

Thus Rawls's account of public reason, like so much else in his later work, responds to difficulties he found in *TJ*'s argument for stability. At the end of this section, I also tried to show how the possibility of a *very fully comprehensive doctrine* is connected—via public knowledge and its limits—to one of the last major changes between *TJ* and *PL* that I said I wanted to explain: Rawls's claim in the second version of "Idea of Overlapping Consensus" that consensus on a single conception of justice is not especially realistic (*PL*, p. 164).

§X.7: Stability, Reflective Equilibrium, and Public Justification

I now want to return very briefly to an objection I raised when I sketched *PL*'s basic stability argument in §X.1. I said then that one of the difficulties with

18. My reading also blunts the force of a seemingly powerful criticism. Michael Perry reads Rawls as saying that citizens should honor the guidelines of public reason so that the outcomes of their deliberations will be legitimate. But, Perry says, to identify the guidelines of public reason just is to show what makes the outcomes of public deliberation legitimate. To assume the principle of legitimacy in order to explain why citizens must honor the guidelines so that their outcomes satisfy that principle is to beg the question of what the right the guidelines are. See Michael Perry, *Religion in Politics: Constitutional and Moral Perspectives* (New York: Oxford University Press, 1999), p. 58.

This is a natural and compelling criticism if we read Rawls's account of public reason, as Larmore and Perry do, as an account of how citizens are "officially [to] decide the basic principles of their association." On my reading, by contrast, Rawls recognizes that citizens rarely make such decisions. He thinks citizens of a WOS are to follow the guidelines of public reason in order to assure one another that they acknowledge the authority of the political conception of justice—including its principle of legitimacy. Assuming that they do acknowledge the authority of the principle is not question-begging. It is just what we would expect Rawls to do if his account of public reason has the purposes I have said it does.

attributing that argument to the Rawls of *PL* is that he seems to show stability using very different arguments than he actually did in that book, especially in the "Reply to Habermas." There, Rawls answers Habermas's question about how an overlapping consensus contributes to stability by presenting various kinds of justification (*PL*, pp. 385ff). Rawls's answer raises questions about the reading of *PL* offered here because none of the kinds of justification Rawls discusses seems to be connected with the "balance of reasons" arguments, the game-theoretic considerations, or the *mutual assurance problem* on which I have said *PL*'s treatment of stability depends. I therefore want to show that Rawls's remarks about justification in the "Reply to Habermas" do not contradict my reading of *PL*'s treatment of stability, but instead give it considerable support.

The "basic case" of justification that Rawls discusses in "Reply to Habermas" is what he calls "public justification" (*PL*, p. 388). Public justification, Rawls says, "happens 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*, p. 387). Rawls implies that public justification "gives the best justification a political conception can have at any given time" (*PL*, p. 388). And so in "Reply to Habermas," Rawls seems to say that a public conception of justice is stable—or most stable—when it is publicly justified. Public justification, as laid out in Rawls's text, seems to be a very different condition of stability than the ones I have identified.

To show that the account of stability as public justification coincides with the account I have attributed to the Rawls of *PL*, I proceed in two steps. First, I return to an idea Rawls introduced in *TJ*, but of which he gives a much more refined treatment in *PL*: reflective equilibrium. I show that the basic stability argument I found in *PL* is an argument for wide and general reflective equilibrium. In the second step, I show that a conception of justice is publicly justified just in case it is in such equilibrium. As before, I focus on justice as fairness of simplicity's sake.

In *TJ*, reflective equilibrium was treated as a state reached by individuals singly (*TJ*, p. 50/44). The same is true of *wide* reflective equilibrium as mentioned in "Reply to Habermas." Thus, suppose that someone takes account of all the arguments of which she is aware for various conceptions of justice and has brought justice as fairness into line with her considered convictions at all levels of generality. Then justice as fairness is in a kind of reflective equilibrium with her convictions and with what she knows of arguments about justice. Endorsement of justice as fairness is what we might call a "reflective equilibrium state" for her. Because of all that she has taken account in reaching it, the reflective equilibrium is wide (*PL*, pp. 384–85, note 16).

By contrast with wide reflective equilibrium, general reflective equilibrium—like general equilibrium in economic theory—is a state reached by all actors taken together. I suggest that the wide and general equilibrium state of a WOS is a state in which:

C_{PL}: Each member of the WOS judges, from the viewpoint of full deliberative rationality, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness.

The suggestion gets some confirmation from the fact that the steps in the basic stability argument for C_{PL} are just the steps that would have to be gone through to show that wide and general reflective equilibrium obtains.

To see this, suppose that everyone in the WOS has reached wide reflective equilibrium, so that acceptance of justice as fairness is a reflective equilibrium state for each. To reach this state, each person has had to take account of the convictions drawn from her comprehensive doctrine. And if everyone has reached wide reflective equilibrium, then *PL's Nash Claim* is true. The basic stability argument got to this conclusion by way of (9.3), which says that everyone's comprehensive doctrine is "either congruent with, or supportive of, or else not in conflict with" justice as fairness and that an overlapping consensus obtains. This must also be true if everyone has reached wide reflective equilibrium.

An equilibrium that is general as well as wide is, Rawls says, "fully intersubjective." To reach it, "each citizen has taken into account the reasoning and arguments of every other citizen" (*PL*, pp. 384–85, note 16). Part of what Rawls means by this is that each person knows that the WOS is a just society, supported by everyone's sense of justice. I believe he also means that each person knows—on the basis of acquaintance with political culture and everyone's adherence to public reason—that everyone else's comprehensive doctrine affirms justice as fairness, and so each knows that an overlapping consensus obtains. Each person can therefore take into account the fact that no one else's conception of the good gives him reason to defect from justice as fairness. This mutual knowledge is what makes it possible for citizens collectively to reach a wide reflective equilibrium that is general. As we saw, it is also what solves the *mutual assurance problem*, so that Rawls can move from *PL's Nash Claim* to C_{PL}. If the desires that tip each person's balance in favor of justice are assumed to be enduring, then justice as fairness will be stable. Indeed, it will be a stable equilibrium state, described by C_{PL}. Thus, given this assumption, the basic stability argument is an argument for what seems to be Rawls's clear implication that a WOS would be stabilized by a wide and general reflective equilibrium.

I have said I want to show that the account of stability I have attributed to Rawls is the same as the account he lays out in the discussion of justification in "Reply to Habermas," and I have said I would do so in two steps. The first step is now complete. To complete the second step, I need to show that a conception of justice is publicly justified just in case its public acceptance is a state of general and wide reflective equilibrium.

As I mentioned a moment ago, public justification is, Rawls says, the "basic case" of justification (*PL*, p. 388). Rawls distinguishes it from what he calls "full justification." Full justification, like the process of reaching wide

equilibrium, "is carried out by an individual citizen" (*PL*, p. 386). Each citizen is assumed to have both a political and a comprehensive view. Full justification concerns the relation between the two. To reach full justification, Rawls says, "the citizen accepts a political conception and fills out its justification by embedding it in some way into the citizen's comprehensive doctrine" (*PL*, p. 386). I take it Rawls means that in order to reach full justification, each citizen either has to show herself that her comprehensive doctrine can take part in an overlapping consensus on justice as fairness, or she knows that it can, and she draws on that knowledge to show herself that she endorses justice as fairness from within her comprehensive view. Either way, full justification depends upon the existence of an overlapping consensus.

Rawls adds rather enigmatically that "Some may consider the political conception fully justified even though it is not accepted by other people. Whether our view is endorsed by them is not given sufficient weight to suspend its full justification in our own eyes" (*PL*, p. 386). I interpret this remark as saying that some members of the WOS will judge, from within their comprehensive view, that they should live up to the demands of justice as fairness unconditionally, while others will judge that they should live up to it only if everyone else does as well. We can sum this up by saying that members of the WOS judge that they should live up to it at least when others will. On this reading, if an overlapping consensus obtains and (9.2) and (9.3) are true, and if each person in the WOS carries out full justification then, just as when each person reaches wide reflective equilibrium, *PL's Nash Claim* is true. Then:

C_N*: Each member of the WOS judges, from within her comprehensive view, that the balance of her reasons tilts in favor of maintaining her desire to live up to the values and ideals of justice as fairness, at least when others live up to those values and ideals as well.

The process of public justification, like that of reaching a wide reflective equilibrium that is general, is carried out by all citizens. In the case of public justification, unlike the case of full justification, "reasonable citizens take one another into account[.]" (*PL*, p. 387) As we saw, this is something each member of the WOS has to do if wide and general reflective equilibrium is to be reached. Rawls completes the thought by saying that in public justification, citizens take one another into account "*as having reasonable comprehensive doctrines that endorse the political conception*" (*PL*, p. 387, emphasis added). This remark implies that public justification of justice as fairness presupposes that it is and is known to be the object of an overlapping consensus. The implication is confirmed a few pages later when Rawls says that that public justification depends upon "the existence and public knowledge of a reasonable overlapping consensus" (*PL*, p. 392). We have seen that general and wide reflective equilibrium depends upon the same things.

How does public justification depend upon "the existence and public knowledge of an overlapping consensus"? Since in reaching public justification, "citizens take one another into account as having reasonable comprehensive

doctrines that endorse the political conception" (*PL*, p. 387), public justification presupposes that everyone carries out full justification. As we just saw, if everyone carries out full justification, then *PL's Nash Claim* is true. Since each person knows that everyone else's comprehensive doctrine "endorse[s] the political conception," each person knows that no one else's view of the good gives him reason to be unjust. This solves the *mutual assurance problem*.

The availability of a solution to that problem explains Rawls's reply to the question "how can public justification of the political conception of justice be carried out?" (*PL*, p. 392). Rawls answers that when an overlapping consensus obtains and is known to obtain, "we hope that citizens will judge (by their comprehensive view) that political values either outweigh or are normally (though not always) ordered prior to whatever nonpolitical values may conflict with them" (*PL*, p. 392). This is just the step I called "C_9" in the basic stability argument. Rawls can move from this step to C_{PL}, the claim that everyone's plan of life makes room for his desire to live up to the values and ideals of justice as fairness. This completes the case for public justification.

I have said that C_{PL} describes the wide and general equilibrium state of a WOS. So showing that justice as fairness enjoys public justification is a matter of drawing "the existence and public knowledge of an overlapping consensus" (*PL*, p. 392) to show that it is that state. As if to sum this line of thought, Rawls says:

> This basic case of public justification is one in which the shared political conception is the common ground and all reasonable citizens taken collectively . . . are held in general and wide reflective equilibrium in affirming the political conception on the basis of their several reasonable comprehensive doctrines. (*PL*, p. 388)

Thus, the basic stability argument I located in *PL* is an argument that a WOS would be in a state of general and wide reflective equilibrium described by C_{PL}. An argument that the WOS would be in such an equilibrium is, in turn, an argument that justice as fairness would be publicly justified in the WOS. The account of stability I have imputed to Rawls and the account he provides in "Reply to Habermas" are therefore essentially the same. While the latter account seems new to that essay, I have shown a number of parallels between the *TJ*'s arguments for stability and *PL*'s basic stability argument. Those parallels, together with the essential identity of the basic stability argument and the treatment of stability in "Reply to Habermas," suggest important similarities between Rawls's earliest treatment of stability and his last.

§X.8: Conclusion

It is time to take stock. In Chapter VIII, we saw why Rawls came to think that *TJ*'s arguments for the stability of justice as fairness failed. We have now seen how Rawls rebuilds the argument by recasting justice as fairness as a political conception of justice.

The Rawls of *PL* thinks that members of the WOS would normally acquire a sense of justice following the process of moral learning laid out in *TJ*, subject to the additions and modifications noted in §IX.2. An important part of that argument is the argument for C_3'. This shows that members of the WOS would acquire desires to live up to the political ideals of justice as fairness. In §§IX.3 and IX.4, we saw what those ideals are: the ideal of full political autonomy, of civic friendship, and the *Ideal of Democratic Governance*. Those ideals are, we saw, specified using principles of right. Members of the WOS can live up to them only by treating the desire to act from principles and values of justice as fairness as regulative of their political life. The treatment of the sense of justice as ideal-dependent is a major development in Rawls's thought. I have tried to show why he developed his thought in that way.

Rawls then uses what I have called the "basic stability argument" to show that members of the WOS would affirm their sense of justice on the basis of their diverse comprehensive doctrines. He assumes that people follow their comprehensive views. The second and third steps of the argument—(9.2) and (9.3)—say that an overlapping consensus would obtain in a WOS. In §X.2 we saw why Rawls thinks it would. To say that an overlapping consensus obtains is not to say that members of the WOS regard every last exercise of power is just or optimal. It *does* require that they regard exercises of power are justified. We saw in §§X.3 and X.4 that Rawls introduced a principle of justification that was new to his view—the liberal principle of legitimacy—to show how it is possible for "citizens of faith" to regard exercises of power this way. Having shown this, Rawls can infer the conclusion I called *PL's Nash Claim*.

Even after reaching that conclusion, Rawls still needs to solve *mutual assurance problem*. He still needs to show, that is, that each member of the WOS needs assurance that everyone else is committed to treating values and ideals of justice as fairness as authoritative. Assurance may be particularly hard to come by if society includes *very fully comprehensive doctrines*, but we have seen how Rawls relies on compliance with the guidelines of public reason to solve the *mutual assurance problem*. Once that problem is solved, Rawls can move to C_9 and C_{PL}, according to which everyone affirms his desire to treat the values of justice as fairness as authoritative—or, more realistically, to C_9' and C_{PL}', according to which everyone affirms the desire to treat values of one or another liberal political conception that way. C_{PL} and C_{PL}' describe equilibrium states. If an overlapping consensus is enduring, each member of the WOS will affirm her sense of justice every time she adopts the viewpoint of her comprehensive doctrine. So the equilibria described by C_{PL} and C_{PL}' are stable, and the WOS will be stably just over time. The argument from (9.2) and (9.3) to C_{PL} and C_{PL}' shows how an overlapping consensus stabilizes. In the previous section, I showed that this account of stability is essentially the same as the one Rawls sketches in "Reply to Habermas."

We saw that in *TJ*, Rawls's arguments for stability depended upon the claim that there was a single set of thin reasons all citizens share—reasons supplied by the desires referred to by C_4a, C_4b, C_4c, and C_4d. Reasonable pluralism opens

the possibility that some people's good may *not* include the satisfaction of the desires and interests that the congruence arguments of *TJ* assume they have. Thus Rawls came to realize, not just that the congruence arguments of *TJ* failed, but that *TJ*'s strategy of demonstrating congruence failed. The strategy failed because it required just institutions to make all conceptions of the good converge in respects that were essential for Rawls's argument.

The existence of an overlapping consensus enables Rawls to reach conclusions about citizens' balances of reasons that are similar to the ones he reached in *TJ*, but to reach those conclusions by a quite different path. For in *PL*, the convergence of desires expressed in (9.1) is merely nominal. If an overlapping consensus obtains, then while every citizen has some comprehensive reasons to affirm her sense of justice, citizens' reasons differ with their comprehensive views, as the disjunction in (9.3) suggests. Crudely put, in *TJ*, there is a single set of desires that everyone has and that gives everyone thin reason to be just; *PL* reverses the order of the quantifiers, showing that each citizen of the WOS has comprehensive reasons to affirm her sense of justice without implying that all citizens' reasons are the same.

Even in *PL*, stability depends upon *some* convergence of conceptions of the good. For the WOS of *PL* is stabilized by the convergence of comprehensive doctrines on justice as fairness, understood now as a political conception of justice rather than a partial comprehensive doctrine. As in *TJ* so in *PL*, the convergence is not fortuitous. It depends upon the forces of social learning at work in the WOS. We saw that in *TJ* and the original *Deweys*, the various desires on which stability depends—including the desires referred to by C_4a, C_4b, C_4c, and C_4d and various ideal-dependent desires—are encouraged by just institutions. In §X.2, I showed how just institutions encourage an overlapping consensus. Thus in *PL* as in *TJ*, justice as fairness, when institutionalized, brings about the convergence that stability requires.

According to the Rawls of *TJ*, the fact that justice as fairness would generate its own support showed that it was inherently stable. As we have seen, the pluralism of the WOS significantly complicates the way justice as fairness generates support for itself. In retrospect, Rawls came to see that the account of stability in *TJ* treated of what he called "the simplest case":

> where the public conception of justice is affirmed as *in itself sufficient* to express values that normally outweigh, given the political context of a constitutional regime, whatever values might oppose them[.][19]

Once he realized that justice as fairness would not be affirmed as "in itself sufficient" and that stability depends upon "the existence and public knowledge of an overlapping consensus" (*PL*, p. 392), he realized that it would not be entirely accurate to say stability *inheres* in justice as fairness. This, in turn, led

19. Rawls, "Political Not Metaphysical," *Collected Papers*, p. 414, note 33 (emphasis added).

the Rawls of *PL* to think that an exhaustive distinction between inherent and imposed stability expresses a false dichotomy.

When there is an overlapping consensus on a liberal political conception of justice—or a family of such conceptions—citizens all accept principles of justice, guidelines of public reason, and a principle of legitimacy that would be acceptable to them as free and equal rational persons. Since these principles and their derivation must be public knowledge, everyone would know that the principles they all agree on would be acceptable to "their common human reason" (*PL*, p. 137). When those principles are institutionalized and commonly acted on, they would elicit a sense of justice. Because moral education would itself be transparent in a WOS, citizens can all see that their sense of justice is not the result of indoctrination or the blind internalization of arbitrary authority. The justice of the WOS is therefore sustained by the fully autonomous activity of its members. Because an overlapping consensus obtains, everyone affirms her sense of justice as part of her good from within her own comprehensive doctrine; her affirmation is itself as willing and transparent as her comprehensive doctrine permits. Stability brought about by overlapping consensus may not be inherent stability. But because it is sustained by fully autonomous activity and by attitudes that are voluntarily affirmed, it is not imposed stability either. Rather, even after Rawls's political turn, it still seems to be stability that is brought about by citizens' acting for what we intuitively regard as reasons of the right rather than the wrong sort. And so it still seems to be—to use Rawls's phrase—stability "for the right reasons" (*PL*, p. xlii).

The fact that Rawls showed how a liberal political conception—or a family of liberal political conceptions—can be stable for the right reasons meant that he did not need to give up the ambition that was evident even in his earliest work. That ambition, as we saw at the end of Chapter II, was to show that human beings can honor just and collectively rational norms over the long run, without an absolute sovereign or a dominant ideology. In both *TJ* and *PL*, showing this requires showing that a liberal political conception can "transform[] of our pattern of final ends" (*TJ*, p. 494/432) so that we acquire and affirm a sense of justice, and that it can do so without violating our freedom. This possibility depends upon showing that our nature allows for the transformation of our ends.

In *PL*, showing this depends upon showing that diverse comprehensive doctrines can develop so that they are "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 for a democratic regime" (*PL*, p. 169). It depends, that is, upon showing that we are the kind of creatures who can reach and sustain an overlapping consensus, at least under the influence of free and just institutions. By showing that our comprehensive views can develop and converge, and therefore that we are creatures of that kind, Rawls hoped to answer Hobbes and the even darker minds of Western thought, whose pessimistic views of human nature are–as we saw

§VII.10—reflected in their claims about how political stability must be maintained. Doing so would carry out the task "Kant gave to philosophy generally: the defense of reasonable faith." In his hands, Rawls hastened to add, "this becomes the defense of reasonable faith in the real possibility of a just constitutional regime."[20] As we shall see, it is also a defense of reasonable faith in the goodness of humanity.

20. Rawls, "Idea of an Overlapping Consensus," *Collected Papers*, p. 448.

XI

Conclusion: Why Political Liberalism?

In this book, I have tried to show why Rawls became dissatisfied with *TJ*'s arguments for stability and why he recast justice as fairness as a political liberalism. I have tried to confirm my interpretation by showing how the many changes between *TJ* and *PL* respond to the difficulties Rawls found in his earlier arguments for stability and congruence.

This book has also been intended as a defense of Rawls's turn to political liberalism. As I indicated in the Introduction, I have tried to defend Rawls's political turn by presenting arguments for it in their most attractive and rigorous form. I believe the problems Rawls found in *TJ*'s treatment of stability are powerful and telling, and I tried to present them as such in Chapter VIII. In my view, Rawls was right to think that he had to remedy those difficulties if he was to show that justice as fairness would be stable for the right reasons. Recasting justice as fairness as a political liberalism was the proper way to remedy them. I have also tried to show why demonstrating stability for the right reasons is important, and I shall return to its significance in the closing section of this chapter.

The kind of defense I have tried to provide stands in sharp contrast to the usual ways of defending philosophical positions. I have not tried to defend Rawls's version of political liberalism by distinguishing it from other political liberalisms on offer and showing that Rawls's is more nuanced or superior. The reason I have not done so is that comparison and defense depend upon the understanding of Rawls's turn to political liberalism that I have tried to provide.

To see this, let us consider Martha Nussbaum's capabilities approach to basic justice, which she says is a form of political liberalism that can be the object of an overlapping consensus.[1] As we saw in Chapter X, Rawls finally concluded that a well-ordered society (WOS) would be characterized by an overlapping consensus on a family of reasonable political conceptions of justice that would contend politically. In saying that her capabilities view would be the object of an overlapping consensus, Nussbaum does not just mean that it can be a member of such a family. Rather, she thinks that her view could be incorporated into a constitution, and so could be the conception that well-orders a liberal society by serving as the public basis for adjudicating at least some of the most important competing claims.[2] Thus, Nussbaum has many of the same ambitions for her conception of justice that Rawls had for his. If we are to assess the relative merits of the two political liberalisms, one of the questions we have to ask is whether Nussbaum's conception of justice would be more or less stable than Rawls's.

Nussbaum defends the conclusion that her conception could be the object of an overlapping consensus by pointing out that it does not presuppose a metaphysical conception of the person that would compete with the conceptions included in fully comprehensive doctrines. Like Rawls, Nussbaum therefore "put[s] no doctrinal obstacles to [her view's] winning allegiance to itself, so that it can be supported by a reasonable and enduring overlapping consensus" (*PL*, p. 40). Rawls insists, however, that the absence of obstacles is a necessary but not a sufficient condition for a view's being a political liberalism. Another condition he imposes on political liberalisms is that they be worked up from ideas implicit in the democratic tradition; some of Nussbaum's critics have asked whether her view satisfies this condition.[3] Now that Rawls's treatment of stability is before us, and we can see what makes an overlapping consensus "enduring," we can see precisely why this question matters.

According to the Rawls of *PL*, the stability of a WOS depends upon citizens' acquisition of various ideal-dependent desires, such as the desire to live as fully autonomous citizens. Stability also depends upon the fact that those desires would be affirmed on reflection. Members of the WOS are, by hypothesis, heirs of the democratic tradition. They think of themselves as free and equal, and so think of themselves as the democratic tradition says they are. One of the reasons they give their desire to live autonomously their reflective endorsement is that they recognize the ideal of full autonomy as an appropriate specification of political freedom.

1. See, for example, Martha Nussbaum, *Women and Human Development*, (Cambridge: Cambridge University Press), p. 5.

2. Nussbaum, *Women and Human Development*, p. 5.

3. See Ruth Abbey, *The Return of Feminist Liberalism*, chapter 12, note 8 (manuscript in progress).

Thus the Rawls of *PL* thinks the stability of justice as fairness depends crucially upon the fact that its political ideals develop and specify ideas found in the tradition of liberal democratic thought. That, I believe, is why he imposes the additional condition on political liberalisms. Nussbaum may be able to argue that her view satisfies that condition. Alternatively, she may be able to argue that it would, if institutionalized, prove stable for the right reasons even if it does not satisfy the condition. Hers is a very sophisticated view and I cannot give it full consideration here.[4] I bring it up now to illustrate two general points. One is that to defend a political liberalism like Nussbaum's, it is not enough to show that the view can be presented as standing free of comprehensive views; it is necessary to ask whether and how that alternative conception of justice would stabilize itself. The other is that defending Rawls's political liberalism by comparing it with political liberalisms put forward by others would have been premature without the detailed understanding of Rawls's treatment of stability that I have tried to provide.

Once Rawls's treatment of stability is before us, we can also see that one objection commonly raised against it is based upon misunderstanding. Critics sometimes suppose that to show an overlapping consensus is possible, Rawls must show that religions in their current state endorse or are likely to endorse the starting premises of justice as fairness, or endorse or are likely to endorse its principles. If it is then asserted that they don't or can't, the argument for the possibility of such a consensus is taken to fail.[5] It should now be clear how badly this objection misses the mark. The idea of an overlapping consensus is introduced at steps (9.2) and (9.3) of *PL*'s basic stability argument to show how justice as fairness, *once institutionalized*, could generate its own support. As we saw in § X.2, Rawls's defense of those steps depends crucially on the formative effect of just institutions. The world in which we live is, however, a world from which the just institutions of a WOS are notably absent. Arguments drawn from the injustice of our world cannot, therefore, tell against the possibility of an overlapping consensus or against the conclusion that justice as fairness would be stable for the right reasons.

It might be objected that an account of stability which depends upon claims that cannot be verified or falsified by facts about our world is an account of little interest. But facts about our world, while not dispositive, are not irrelevant to the possibility of an overlapping consensus, as I tried to show in §X.2. I have also tried to show how questions about the realistic possibility of an overlapping consensus bear on the view we have of our own nature. Questions

4. I examine it in my "Claims and Capabilities," forthcoming in the *Library of Living Philosophers* volume on Martha Nussbaum; available on request.

5. Cf. Barry, "Search for Stability," pp. 910ff. For similar criticisms, see John Gray, *Two Faces of Liberalism* (New York: The New Press, 2000), pp. 23–25; George Klosko, "Rawls's 'Political Philosophy' and American Democracy," *The American Political Science Review* 87 (1993): pp. 348–59; L. Gregory Jones, "Should Christians Affirm Rawls's Justice as Fairness? A Reponse to Professor Beckley," *Journal of Religious Ethics* 16 (1988): pp. 251–71, p. 260.

about how we view our own nature do not lose their interest simply because we do not live in a just society at the moment, nor because the disordered state of our world makes it difficult to see how we could make the transition to a society which is well-ordered.

There is one set of concerns about political liberalism that I do want to address. Those concerns stem from the fact that political liberalism founds justice as fairness on ideas drawn from liberal democratic culture, and works out their implications. Because the principles are derived from starting points for which Rawls does not attempt to provide an intellectual foundation, he seems to be left without any philosophical grounds for criticizing those who are prepared to deny the starting points as well. If some society denies that citizens ought to be accorded rights and liberties because it denies that its citizens are to be treated as free and equal, then—so the objection goes—Rawls is left without any philosophical objection to make. Moreover, if a society without a democratic past is in transition, as many societies were after the collapse of European communism, political liberalism cannot provide that society or its democratic activists with the intellectual resources needed to defend liberalization. Finally, it is said, even those who endorse the claims from which the later Rawls begins may find political liberalism unsatisfying, since it does not vindicate those starting points philosophically. Rawls reaches conclusions that make powerful normative demands, but he starts from what we citizens of Western liberal democracies simply happen to believe.

These objections all charge that political liberalism fails to provide the principles of justice the right kind of authority. The Pivotal Argument for the principles of justice that I introduced in Chapter I has proven to be a powerful analytic device. It helped to articulate the *Public Basis View* of Rawls's transition to political liberalism. Some of its premises played an important role in the congruence arguments of *TJ*, and public knowledge of some of its premises are instrumental in citizens' development of a sense of justice. The Pivotal Argument also makes it possible to pinpoint the grounds of the charge against political liberalism that I now want to consider. Furthermore, it enables us to see both exactly where sympathetic readers have tried to rebut that charge by providing foundations for justice as fairness, and to see just why those attempts are mistaken. Once we see where they go wrong, we will be better positioned to see how the charge should be answered.

§XI.1: The Moral Basis of Political Liberalism?

Recall that in a WOS, the Pivotal Argument provides a public argument, or a public basis, for the principles of justice. We saw in §I.2 that in the WOS of *TJ*, that argument ultimately derives the principles of justice from the claim that persons are naturally free and equal. We saw in §I.4 that the turn to political liberalism required a reformulation of the Pivotal Argument. In the WOS of

political liberalism, the public basis of the principles would be, or would include, the argument that results from replacing some of the original premises of the Pivotal Argument with what I have called their "political analogues." These are premises in which "citizens" is substituted for "persons."

Thus, the original version of the Pivotal Argument begins with:

(1.1) We are free and equal *persons* who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

The reformulated version begins with:

(1.1') We are free and equal *citizens* who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

Again substituting "citizens" for "persons," the reformulated version appeals to:

(1.5') Our society respects us as the kind of *citizens* (1.1') says we are only if the principles governing the ways the basic structure of our society distributes primary goods are acceptable to us as such *citizens*.

It moves from (1.5') to:

(1.6') Principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal *citizens*.

The argument then moves from (1.6'), via a distinctive understanding of "free and equal citizens," to:

(1.8') The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal *citizens* is the decisive determining element of the choice.

Since the choice situation referred to in (1.8') is said in (1.9') to be the original position, the argument concludes that distribution must be governed by Rawls's principles of justice.

According to political liberalism, then, the principles of justice are ultimately conditional on the liberal democratic view of citizenship expressed in (1.1'). But because the Rawls of *PL* treats that view as a starting point, he seems unwilling to say what reasons we who endorse (1.1') have for endorsing it, and why those who do not endorse it can be rationally criticized for not doing so. Put somewhat more crudely, the Rawls of *PL* seems unwilling to say what gives the view of citizenship expressed in (1.1') its normative force or purchase.

The revised version of the Pivotal Argument shows that Rawls's unwillingness to do so does not just matter for the acceptability of his starting point

(1.1'). It also affects his ability to get from (1.1') to the conclusion that distribution of primary goods must be governed by the principles of justice. For Rawls can reach that conclusion only if he can move from (1.5') to (1.6'), and that move seems to depend upon the assumption that free and equal citizens must be respected as such. But unless Rawls explains the normative force of the view of citizenship expressed by (1.1'), it will be unclear what it is about citizenship that demands respect. In that case, it will not be at all clear how political liberalism can provide the principles of justice the authority Rawls wants them to have.

Charles Larmore developed his own version of political liberalism well before Rawls published *PL*.[6] He has also written penetratingly about the historical developments that made comprehensive liberalism controversial and about Rawls's political turn. Larmore argues that Rawls moved to political liberalism because he thought that principles which are to be enforced with public power, such as the principles of justice, are subject to what we might call an "acceptability requirement." Rawls thought, Larmore writes, that "basic political principles should be suitably acceptable to those they are to bind."[7] Rawls moved to political liberalism because he recognized that founding the principles on comprehensive liberalism would violate this requirement. According to Larmore, Rawls's transition to political liberalism therefore raises the question of what grounds the acceptability requirement. Larmore argues that the requirement is derived from a moral principle enjoining respect for persons. Principles that are to be coercively enforced must be "suitably acceptable" because if they are not, those who are coerced to comply with them are not respected as persons with reason and will. Because the principle enjoining respect for persons grounds the acceptability requirement, it expresses the "moral core of liberal thought"[8] and forms the basis of Rawls's political liberalism.

Though he does not put it this way, Larmore's attempt to identify the basis of political liberalism can be read as an attempt to help Rawls answer the objections I have said he seems to face. Since the "basic political principles" to which Larmore refers include the principles by which the basic structure is to distribute primary goods, the acceptability requirement—as applied to these principles—just is the step in the revised Pivotal Argument that I called (1.6'). The step from which the argument moves to (1.6') says, in effect, that satisfying the acceptability requirement is a condition of respecting citizens. If we think of citizens as persons considered in a certain role, namely the role of those who are subject to coercion by public power, then we can see why Lamore thinks a principle that enjoins respect for them is what Rawls relies on to move from (1.5') to (1.6'). If that move were justified, Rawls could get the conclusion

6. Charles Larmore, "Political Liberalism."
7. Larmore, *Autonomy of Morality*, p. 146.
8. Larmore, *Autonomy of Morality*, p. 143.

about the principles of justice that he wants. The authority of the principles would then depend upon the authority of the moral requirement that persons be respected as such.

The claim that political liberalism depends upon a Kantian or quasi-Kantian requirement of respect for persons is, I believe, widely held. Larmore states it especially clearly, and I have used the Pivotal Argument to make his version of the claim precise. Unfortunately, the solution Larmore offers Rawls comes with a very high price, for the requirement of respect for persons is not one that Rawls can account for from within political liberalism.[9] Larmore thinks that Rawls's attempt to frame a political conception of justice that is freestanding (*PL*, p. 10) is therefore, at best, a qualified success.

We can see just why Larmore thinks Rawls had to pay so high a price by unpacking a very interesting exegetical claim on which Larmore's argument seems to depend.

Larmore says Rawls "came to call" the acceptability requirement instantiated by (1.6') the principle of legitimacy.[10] Since (1.6') imposes a condition on the choice of principles of justice, Larmore must think that the principle of legitimacy imposes a condition on the adoption of the principles and not just, as I argued in §§X.3 and X.4, on their implementation.

This interpretation derives some support from Rawls's attempt to develop the contract tradition. The wording of the principle of legitimacy makes clear that it imposes a condition on the exercise of power. If Larmore thinks that that principle governs the adoption of the two principles, it must be because he thinks the adoption of principles of justice is itself an exercise of power. As I said in §§X.3 and X.4, I read the principle of legitimacy as applying to exercises of the two kinds of power that have been recognized by contract theorists at least since Locke: constituent power, exercised at the second stage of *TJ*'s four-stage sequence, and ordinary power, exercised at later stages. But Rawls says that he "carr[ies]" social contract theory "to a higher order of abstraction" (*TJ*, p. viii/xviii) than Locke did, in effect by introducing a stage prior to the constitutional stage (see *TJ*, p. 11/10). Though Larmore does not defend his interpretation this way, we can read him as suggesting that by introducing a contract prior to Locke's, Rawls recognizes an additional kind of power, exercised in choice of principles for the basic structure, which is also subject to the principle of legitimacy.

But interpreting Rawls this way requires reading him as making a serious mistake, a mistake that can be stated precisely with the help of the Pivotal Argument. Rawls introduces the original position to identify principles of justice that are acceptable to us as free and equal, and so to enforce the acceptability requirement instantiated in (1.6'). That is why the Pivotal Argument moves from the sixth step to its conclusion by way of:

9. Larmore, *Autonomy of Morality*, pp. 140, 150.
10. Larmore, *Autonomy of Morality*, p. 146.

(1.10) The principles governing the ways the basic structure distributes primary goods must be acceptable in the OP.

If Larmore is correct in thinking that (1.6') instantiates—and hence that (1.10) depends upon—the principle of legitimacy, then the function of the original position would really be to help us identify the "basic political principles" that can legitimately be enforced. But as we saw in Chapter X, Rawls thinks that principle of legitimacy—the principle that enjoins us to look for basic principles that can legitimately be enforced—gets its normative force from the fact that it would be *adopted* in the original position. So if Rawls introduces the original position to enforce to (1.6'), as I have said, and if he thinks that (1.6') instantiates the principle of legitimacy, as Larmore contends, then his account of the principle's normative force would be patently circular. Larmore does seem to think Rawls argues in a circle, and he thinks Rawls can break out of the circle only by appealing to a requirement for which political liberalism does not itself try to account.[11]

It is important that a strategy which might seem to eliminate the circle cannot, in fact, salvage Rawls's argument. In §VII.9, we saw that it is possible to defend the two principles of justice without appealing to the original position. It is therefore possible to move from (1.6') to the conclusion of the Pivotal Argument without going by way of (1.10). This possibility shows that the normative force of the principles does not really depend upon their adoption in the OP after all. Seeing that it does not, we may try to remove the circle by using the OP to justify the principle of legitimacy, and hence (1.6'), and we may then move to the principles without relying on the OP. But if Larmore is correct, the difficulty with Rawls's reasoning does not stem from his reliance on the OP. It really arises from his claim that "the argument . . . for the principle of legitimacy is much the same as . . . the argument for the principle of justice themselves" (*PL*, p. 225). For the arguments for the principles of justice—and, according to the passage I just quoted, for the principle of legitimacy—depend upon showing that the principles are acceptable to us all as free and equal citizens, just as (1.6') says. But on Larmore's reading, the principle of legitimacy is what requires us to look for mutually justifiable principles of justice. So long as the principle of legitimacy and the principles of justice are alleged to have the same basis, he thinks Rawls is stuck in the circle.

I do not read Rawls as offering a circular argument because, for reasons I spelled out in §§X.3 and X.4, I take the principle of legitimacy to guide the application of principles of justice but not their adoption. If my reading of the principle of legitimacy is correct, then Rawls does not appeal to it to justify (1.6'). But in that case, Larmore would insist, Rawls needs to provide some other justification for it, and a principle of respect for persons is the only plausible candidate. Moreover, Larmore's argument—like the objection I said

11. Larmore, *Autonomy of Morality*, p. 151.

Rawls seems to face—shows that Rawls must acknowledge some principle of right that is more basic than the principles of justice, the moral force of which *grounds* rather than *depends upon* its acceptability to us as free and equal. That is Larmore's real point, and it is why he thinks political liberalism must have a moral foundation. My counterargument establishes what seems to be the merely verbal point that that foundation is not expressed by the principle of legitimacy.

Larmore says his claim that political liberalism depends upon an imperative of respect for persons is similar to Ronald Dworkin's claim that justice as fairness is founded on each person's right to equal concern and respect.[12] The Pivotal Argument enables us to see just how similar they are. I have said that Larmore appeals to the imperative of respect for persons to justify step (1.6') of the revised version of the argument. When I discussed the original version of the Pivotal Argument in §I.3, I said that the move to the sixth step of that version—to (1.6) rather than to its political analogue (1.6')—seems to be justified by appeal to Dworkin's right to equal concern and respect. Larmore thinks the imperative of respect is, in effect, enforced by the OP, since the imperative of respect imposes an acceptability requirement on political principles and, according to (1.10), principles satisfy that requirement only if they would be adopted in the OP. As we saw in §I.3, Dworkin thinks that "the original position is well designed to enforce the abstract right to equal concern and respect."[13] Because both Larmore and Dworkin think Rawls used the OP to enforce a prior norm, both think Rawls argued "through" the original position.[14]

I introduced Larmore's reading because it seemed to promise an answer to the objection I raised at the end of the last section—the charge that the Rawls of *PL* cannot account for the authority of the principles of justice. Larmore thinks he can account for their authority only by appealing to the requirement of respect for persons. This answer, I said, comes with a high price. It is, however, no higher than the price that Dworkin exacts, for Dworkin seems to have thought that the norm of equal concern and respect is what ultimately grounds the authority the Rawls of *TJ* takes the principles of justice to have. Thus, both Larmore and Dworkin think Rawls must appeal to a norm that is prior to principles of justice and that has a different source than they do, a norm for which Rawls's theory cannot account. We can, I believe, see how Rawls would respond to this common objection to political liberalism—and to Larmore—by spelling out his reply to Dworkin.

12. Larmore, *Autonomy of Morality*, p. 148, note 18; see Dworkin, "The Original Position," p. 181.

13. Dworkin, "The Original Position," p. 181.

14. Larmore, *Autonomy of Morality*, p. 76; Dworkin, "The Original Position."

§XI.2: A Conception-Based View

Because Dworkin reads Rawls as grounding the principles on a right to equal concern and respect, he takes Rawls's view to be "rights-based." In response to Dworkin's reading, Rawls said that "justice as fairness is a conception-based or an ideal-based view."[15] The context of this remark leaves Rawls's meaning somewhat obscure, and that is why Larmore observes that Rawls "never adequately responded" to Dworkin.[16] We can, however, begin to grasp what Rawls had in mind by looking again at the inference Dworkin thinks is justified by the right to equal concern and respect.

Recall that the first step of the Pivotal Argument in its original version is:

(1.1) We are free and equal persons who can reflect upon the ends we pursue, and can assess social arrangements in light of our own interests and ends.

As we have seen, the right to equal concern and respect can be used to license the move from:

(1.5) Our society respects us as the kind of persons (1.1) says we are only if the principles governing the ways the basic structure of our society distributes primary goods are acceptable to us as such persons.

to:

(1.6) The principles governing the ways the basic structure distributes primary goods must be acceptable to us as free and equal persons.

An appeal to Dworkin's right is not the only way to license the inference. Suppose that the Rawls of *TJ* thought:

(1.5.1) Our society is a fair scheme of social cooperation among free and equal persons only if it respects us as the kind of persons (1.1) says we are.

(1.5) and (1.5.1) imply:

(1.5.2) Society is a fair scheme of social cooperation among free and equal persons only if the principles governing the ways the basic structure distributes primary goods are acceptable to us as such persons.

Now suppose Rawls also thought that:

(1.5.3) Society should be a fair scheme of social cooperation among free and equal persons.

15. Rawls, "Political Not Metaphysical," *Collected Papers*, pp. 400–401, note 19.
16. Larmore, *Autonomy of Morality*, p. 76, note 13.

Then Rawls could make the move he wants, since (1.5.2) and (1.5.3) imply (1.6).

The suggestion that this is how Rawls would get to (1.6) derives some support from the way the new steps fit into the larger sweep of the Pivotal Argument, and from the role that the original position can then be seen to play.

Recall that the Pivotal Argument moves from (1.6), via a claim about freedom and equality, to

(1.8) The principles governing the ways the basic structure distributes primary goods must be acceptable in a choice situation in which our nature as free and equal persons is the decisive determining element of the choice.

We have seen that the Rawls of *TJ* constructed the original position so that:

(1.9) The OP is a choice situation in which our nature is the decisive determining element.

(1.8) and (1.9) imply that

(1.10) The principles governing the ways the basic structure distributes primary goods must be acceptable in the OP.

Because Rawls argues that his two principles would be chosen in preference to other principles in the OP, he concludes that:

C_1: The distribution of primary goods by the basic structure must be governed by the two principles.

In §III.2, I noted that the first step in the Pivotal Argument, (1.1), expresses a partial conception of the person. Step (1.5.1) asserts a connection between that conception and a partial conception of society, the conception of society as a fair cooperative scheme. Step (1.5.3) then says that that conception of society is to be realized. Thus, the move from (1.5) to (1.6) shows that the acceptability requirement expressed by (1.6) is conditional on the desirability or importance of realizing that conception.

If that conception of society is to be realized, fair terms of cooperation need to be identified. The movement from (1.6) to (1.10) shows why they are to be identified using the OP. Once those terms are identified, Rawls can specify the conception of a free and equal person into the ideal of full autonomy by reference to the principles by which such persons regulate their conduct. And he can specify the conception of society into the ideal of the WOS of justice as fairness by reference to the principles that regulate its basic institutions. Thus, if we suppose that Rawls moves from (1.5) to (1.6) as I have said he does, the structure of the Pivotal Argument suggests that the OP is introduced to show how the conceptions of the person and society introduced at (1.1) and (1.5.1) are to be further specified into ideals that are connected and jointly realized.

Not surprisingly, this is precisely the role Rawls says the OP is introduced to play in the original *Deweys*. There Rawls says that "justice as fairness tries to uncover the fundamental ideas (latent in common sense) of freedom and equality, of ideal social cooperation, and of the person, by formulating what I call 'model-conceptions.'"[17] He continues "the two basic model-conceptions of justice as fairness are those of a *well-ordered society* and of a *moral person*."[18] Since in justice as fairness, moral persons are free and equal rational persons, the second of these model-conceptions is the conception of the person I have said is introduced in (1.1). That conception, together with the model-conception of a WOS, "depict certain general features of what a society would look like if its members publicly viewed themselves and their social ties with one another in a certain way."[19] And since in justice as fairness, a WOS is cooperative and fair, I take the first model-conception to be the conception of society introduced at (1.5.1), that of a fair scheme of social cooperation.[20] Rawls then confirms that the original position does indeed play the role it would have to play if my reading of the move from (1.5) to (1.6) is right. He says that the OP "is a third and mediating model-conception between the model-conception of the moral person"—introduced at (1.1)—"and the principles of justice that characterize the relations of citizens in the model-conception of a well-ordered society"—introduced at (1.5.1).[21] Thus on my reading as on Dworkin's, Rawls argues through the OP to the principles of justice. But on my reading *unlike* Dworkin's, he argues through it from model-conceptions of the person and of society rather than from a right to equal concern and respect. That is what he had in mind in calling his view "conception-based" rather than "rights-based."

We have seen that Rawls's turn to political liberalism meant recasting the Pivotal Argument so that the conception of the person introduced at the first step becomes a conception of the citizen, and the fifth and sixth steps are replaced by their political analogues. Steps (1.5.1), (1.5.2), and (1.5.3) also have political analogues, which result by replacing "person" in them with "citizen." The political analogue of (1.5.1) is:

(1.5.1') Our society is a fair scheme of social cooperation among free and equal citizens only if it respects us as the kind of citizens (1.1') says we are.

17. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 307.

18. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 308 (emphasis original).

19. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 308.

20. Indeed at "Domain of the Political," *Collected Papers*, p. 480, Rawls seems to take the two basic intuitive ideas as those of the moral person and of society as a fair cooperative scheme; the WOS drops out of the picture as a basic idea.

21. Rawls, "Kantian Constructivism in Moral Theory," *Collected Papers*, p. 308.

This step, together with (1.5'), implies:

(1.5.2') Society is a fair scheme of social cooperation among free and equal citizens only if the principles governing the ways the basic structure distributes primary goods are acceptable to us as such citizens.

The political analogue of (1.5.3) is:

(1.5.3') Society should be a fair scheme of social cooperation among free and equal citizens.

From (1.5.2') and (1.5.3'), Rawls can get to (1.6'). We saw that reading (1.5.1) and (1.5.3) into the Pivotal Argument enabled Rawls to reach (1.5.2) and (1.6) without appealing to a right to equal concern and respect. Reading their political analogues into the revised version enables him to reach (1.6') without appealing to the principle that Larmore takes to be the basis of his view, the principle enjoining respect for persons. And so justice as fairness is conception-based rather than what we might call "respect-based."

Of course, justice as fairness would still be respect-based if Rawls thought that society is fair only if citizens are respected as free and equal, as (1.5.1') says, and he thought that society should be fair, as (1.5.3'), because he thought that societies are obliged to respect their citizens. Reading Rawls this way assumes that the notion of respect in (1.5') and the steps leading up to it is a normative notion that is given prior to the principles of justice and is capable of doing independent work. But that assumption is mistaken. The conception introduced in (1.5.1') is a conception of a society in which members conduct themselves as, and are treated as, free and equal moral citizens. The task of a theory of justice is that of identifying the principles that are appropriate for regulating a society of citizens who are conceived of in this way, who think of themselves in this way and who want to live as free equals.

Having looked at the Kantian Interpretation of justice as fairness in Chapter VII, we are now in a position to see how Rawls does that without relying on a prior requirement of respect. Recall that when we looked at the Kantian Interpretation, we saw that the Rawls of *TJ* thought we can conduct ourselves as free and equal moral persons by acting "on the principles that would be chosen if this nature [as such persons] were the decisive determining element" (*TJ*, p. 253/222). This is what I called the *KI Claim*. In §VII.5, we saw that the grounds of that claim are straightforward. Crudely put, we live as rational beings when our thought and action are governed by principles the content of which is determined by our nature as rational creatures rather than by, say, our appetites. We live as free beings when the content of those principles is determined by our freedom. We live as equals when the content of those principles is determined by our equality.

As I indicated in §IX.3, the Rawls of *PL* relies on the political analogue of that claim, a version of the claim that refers to "citizens" rather than "persons." And so he thinks that citizens live together as free equals when their relations

are regulated by principles that would be chosen if their choice were determined by their freedom and equality. Their society treats them as free equals when its basic institutions are regulated by those principles. The original and revised versions of the Pivotal Argument Step express this claim in the third, fourth, and fifth steps by saying that society "respects" them as free equals when it is regulated by those principles, but the notion of respect does not carry independent moral weight. Rather, 'respects' in those steps is roughly synonymous with 'treats.' The warrant for the third and fourth steps in the revised version of the argument—and hence of (1.5')—is the political analogue of the *KI Claim*.

The political analogue of the *KI Claim* directs us to look at the outcome of a choice situation to identify the principles which would regulate a society of free equals, a choice situation which (1.9') identifies as the OP. We saw that interpolation of (1.5.1) and (1.5.3) into the original version of the Pivotal Argument suggested that the Rawls of *TJ* introduced the OP to show how the model-conceptions of the person and society were to be specified into ideals and simultaneously realized. We also saw that suggestion confirmed in the original *Deweys*. We can now see that interpolation of the political analogues of these steps into the revised version suggests that the Rawls of *PL* assigns the OP a similar role. That suggestion receives less definitive confirmation in the revised *Deweys*, though Rawls does write there that "we introduce an idea like the original position because there seems to be no better way to elaborate a political conception of justice for the basic structure from the fundamental idea of society as an ongoing and fair system of cooperation between citizens regarded as free and equal" (*PL*, p. 26). But just as it is a mistake to think that the role of the OP is to "enforce" a right to equal concern and respect, so it is a mistake to think that its role is to enforce a prior requirement of respect. Crudely put, the political analogue of the *KI Claim* enables Rawls to rely on a procedural interpretation of "respect" that accomplishes what a substantive interpretation might have done. By appealing to a procedural rather than a substantive interpretation, Rawls avoids appealing to a requirement that is independent of the principles. Thus, to paraphrase Rawls's reply to Dworkin, "the force of the [requirement of] respect, is covered in other ways."[22]

§XI.3: Defending Political Liberalism

Rawls's reliance on the political analogue of the *KI Claim* may be surprising, and may seem to imply that Rawls relies on a Kantian justification for the principles of justice after all. It would, however, be a mistake to draw this conclusion. The *KI Claim* is a conditional. It implies that *if* we are to live as free and equal citizens, we must govern our conduct by principles the content of

22. Rawls, "Political Not Metaphysical," *Collected Papers*, pp. 400–401, note 19.

which is determined by the nature of our citizenship. It does not assert, unconditionally, that it is important or valuable to live free or fully autonomous lives, nor does it assert that we are required to conduct ourselves as fully autonomous citizens.

Clearly Rawls thinks we should conduct ourselves as the kind of citizens (1.1') says we are. Moreover, in relying on (1.5.3'), Rawls supposes that society "should" be a fair scheme of social cooperation. I have claimed that the "should" of (1.5.3') does not depend upon a prior imperative of respect for citizens or upon citizens' prior right to equal concern and respect. Rather, we to whom justice as fairness is addressed are assumed to have the conception of citizenship expressed by (1.1') and the conception of our society as a fair scheme of social cooperation. The "should" of (1.5.3') expresses the way society must be *if* we are to satisfy the interest we have in understanding and living up to this view of ourselves and our society.

Since the Pivotal Argument relies on (1.1') and (1.5.3') to reach the principles of justice, this claim brings us back to the objection to which I said Larmore's interpretation may have seemed to promise an answer. Isn't the Rawls of *PL* simply spelling out the implications of ideas that people who grew up in a democratic society happen to have and happen to want to live up to? If so, what normative authority do the principles of justice have once Rawls makes the turn to political liberalism? These questions can be asked from the point of view of members of the WOS, but the more pointed and troubling form of the question is clearly the question that arises for Rawls's readers in the world as it is. What authority do his principles have for those of us who live in liberal democracies as they are? What, if anything, does political liberalism have to offer to societies that are not democratic, and in which Rawls's starting points are not widely shared?

Part of answering these questions consists in showing that the objection which gives rise to them is badly posed. It is a mistake to imply, as the objection does, that our interests in living as free equals and in a society which is fair are interests we just happen to have. The interest we take in being or living as free and equal citizens, or in participating in a fair cooperative scheme, is not a desire that arises spontaneously like the desire for sleep, nor is it an arbitrary desire we lack reasons to satisfy. These claims are evident, I think, when we think about how we acquire these desires and what their objects are.

The interests we take in living up to views of ourselves or our society are, of course, conception-dependent. The conceptions on which they depend are ones we encounter in the justification of political outcomes in a liberal democracy. A specific scheme of liberties, for example, may be justified as a set of arrangements that allows us to live freely. A distributional scheme or a tax provision may be justified on the ground that it is egalitarian or fair. If we approve of the scheme of liberties or the distribution of benefits and burdens, we may then acquire the desire to live up to a view of ourselves or our society which is appealed to to justify the arrangement. If we are firmly convinced that those arrangements are just, the judgment that they are can then be offered as a

reason to affirm the conception of freedom or fairness that is used to justify it. We can cite as a reason for wanting to live freely or fairly that living freely or fairly requires living under a certain scheme of liberties or a distributive scheme that strikes us as just.

The connection between abstract conceptions of freedom, equality, and fairness on the one hand, and political outcomes on the other, suggests a second way in which conception-dependent desires can be acquired. We may take an interest in various liberties—liberties of speech, association, and conscience, for example. If we also affirm an interest in living as a free citizen, it may be because we have abstracted and generalized from what we take these cases to have in common. We may avow an interest in living as free citizens because we know that we have interests in the instances of freedom which are most salient in our political environment. If we have considered judgments that we ought to be free in these ways, then the particular judgments on which the abstract generalization is based can be used to justify it.

The interests we take in living up to democratic conceptions of citizenship and society are acquired interests, and the circumstances in which we acquire them provide us reasons for satisfying them. Since we have reasons to satisfy them, we have reasons to accept Rawls's principles. Conversely, the view of ourselves as free and equal and the claim that our society ought to be fair are ones we could give up only at the very great cost of giving up the more particular judgments of justice on which they are based.

The principles can also be defended by pointing to the deeper understanding they give us of the basic democratic ideas from which Rawls begins.

From Chapter III onward, I have stressed that Rawls's processes of theory construction, and of identifying the principles of justice, add considerable refinement and depth to the liberal democratic conceptions of citizenship and society. For example, in developing a conception of justice that appropriately regulates political relations among citizens who think of themselves as free and equal, Rawls distinguishes various kinds of freedom and uses the principles to give content to the notion of full autonomy. He also develops our pre-theoretical conception of practical rationality quite considerably. He stresses that part of our capacity for practical reasoning is a capacity to justify our conduct to others, and he brings to prominence the moral interest that he supposes we take in developing and exercising that capacity. In particular, he specifies our democratic conception of citizenship by imputing to us an interest in conducting ourselves in ways we can justify to others and in cooperating with them on fair terms. The problem of identifying principles of justice can then be posed as the problem of identifying principles for the basic structure that enable us to conduct ourselves in these ways.[23] Once the basic problem has been made more tractable, and the principles have been defended, we then have before us ideals of conduct and of society that tell us

23. In this brief discussion of practical reasoning, I am indebted to Freeman, "Reason and Agreement."

considerably more than we previously knew about what it is to live as free and equal rational citizens in a society that is fair. Justice as fairness therefore draws on the principles to provide an account of what the desires to live as a free and equal rational citizen and to take part in a fair cooperative scheme are desires *for*. If that account strikes us as correct or reasonable, then its doing so tells in favor of the principles.

We who live in liberal democratic societies think of ourselves as being, in some way, free and equal citizens. By deepening our understanding of the desire to live as such citizens, justice as fairness offers to deepen our understanding of the view we take of ourselves. In the revised *Deweys*, Rawls said that when justice as fairness is fully publicized in the WOS, citizens "are presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain" (*PL*, p. 71). The same is true of us, Rawls's readers, when we come to understand justice as fairness. If the conclusions about our citizenship to which justice as fairness leads are conclusions we accept, then the fact that justice as fairness leads us to them gives us reason to endorse it. If those conclusions are endorsed by our comprehensive doctrines, then we have still more reasons.

Rawls's conception of justice is also supported by considerations of system. Justice as fairness is a theoretical construction that explains judgments in which we have confidence by identifying principles that would lead a person with a sense of justice to arrive at them. The second principle yields answers to questions about which we may initially have been uncertain, such as questions of economic justice. Because the argument for both the principles draws on the same conceptions of ourselves and of society, the answers Rawls proposes have a unified rationale.[24] Furthermore, Rawls was surely right that the democratic tradition is at odds with itself about the relative claims of liberty and equality (*PL*, p. 4). By drawing on basic democratic ideas to reconcile those competing claims, the principles bring greater coherence to that tradition of thought. Rawls's theory is also elegant and parsimonious, introducing simplifying assumptions to isolate the fundamental problems the theory has to address.[25] The systematic virtues of explanatory power and scope, clarity, rigor, and unity are arguably sources of justification when they are exhibited by scientific theories;[26] perhaps the same is true of theories of justice.

Finally, when recast as a political liberalism, justice as fairness satisfies three other important substantive, rather than formal or systematic, desiderata of a public conception of justice. It is capable of providing reasoned

24. On this point, see Cohen, "Democratic Equality," p. 729.

25. I take Rawls's assumption that members of the WOS are fully cooperating members over a complete life to be a simplifying assumption; I discuss this in some detail in "Claims and Capabilities."

26. Ernan McMullin, "Values in Science," *PSA* 1982, vol. 2, ed. Asquith and Nickles, pp. 3–28, especially pp. 14ff.

answers to a wide range of fundamental political questions. As we saw in §§X.3 and X.4, its accounts of public reason and legitimacy allow for the kind of good faith disagreement about those questions that characterizes politics as we know it. And as we saw in Chapters IX and X, if institutionalized and publicized, it could stabilize itself, even if agreement on a family of liberal political conceptions is more likely. These considerations give us further grounds for endorsing justice as fairness.

I have tried to answer the objection that justice as fairness lacks authority for us because it begins from basic ideas found in democratic culture. I have argued that our endorsement of them depends, in part, upon the consequences to which they lead, and upon seeing how the principles of justice specify those conceptions into ideals which are worthy of our aspiration. As these arguments suggest, the problem with the objection—and with attempts to answer it by locating a foundation for political liberalism—is the failure fully to appreciate the justificatory role of reflective equilibrium. We may judge that citizens should be respected by their society, or that they have a right to equal concern and respect, but these judgments do not have any special status. As I have tried to show, the best reconstruction of Rawls's arguments may try to explain those judgments in other terms.

Of course we, Rawls's readers, do not have all the same grounds for affirming justice as fairness that members of the WOS would have. We may find that justice as fairness is in wide reflective equilibrium for us, but since others do not endorse it and since it does not well-order our society, it is not in general reflective equilibrium. We do know, though, that if justice as fairness were adopted, we would then be in the position of having more reasons to endorse it. We can cite as reasons for endorsing justice as fairness that if it well-ordered our society, we would then be in the position to realize certain goods—such as the goods of realizing full autonomy and the *Ideal of Democratic Governance*—which are attractive to us in prospect and which we now think would weigh very heavily with us in a WOS, given the kind of persons we can hope to be in a just society.

Reflective equilibrium may provide an unsatisfying answer to the objection I have tried to address. We can only reach reflective equilibrium by ascertaining the relative weights of our judgments. Many of those judgments themselves either are or reflect further judgments about the weights and balances of reasons. We attach very great weight to freedom of religion, for example, because we think that the goods of living in a religiously free society far outweigh the goods that could be had in a society in which religious freedom is curtailed. If someone rejects the view of citizenship expressed in (1.1'), and the conviction expressed in (1.5.3') that our society should be fair, we can only bring forward other considered judgments to show the costs of the rejection. If the weights he attaches to democratic values are consistently different than ours, so that he is willing to pay the costs of his rejection, there is nothing else to be said.

The disposition consistently to weight values in one way rather than another is the building block of a character.[27] If the basic intuitive ideas from which Rawls begins are in reflective equilibrium with our considered judgments of value, it is because of the character we have. If we find it too costly to give up those ideas, it is because of the kind of people we already are. When we looked at *TJ*'s arguments for congruence and *PL*'s basic stability argument, we saw that members of the WOS would not regret being the kind of persons they are because of the way their balances of reasons tilt, and that their balances of reasons tilt as they do because of the way the institutions of the WOS have shaped their characters and their comprehensive doctrines. We have to recognize that our balances of reasons tilt as they do because of the way our characters and comprehensive doctrines have been shaped by our liberal democratic institutions.

Justice as fairness will have some appeal even in societies without well-developed liberal democratic cultures if members of those societies share some of our considered judgments. It is sometimes lamented that it has less to say to these societies than we might like.[28] This should not be surprising, since liberal democracy is not just a system of government or even a kind of sovereignty,[29] though it is both of those things. It is also a way of life that forms our views of ourselves and—as we shall see in the next section—of the world. That is why, as we have already seen, a liberal conception of justice can be stable for the right reasons once it is in place. Liberal and democratic institutions may have to emerge first as a modus vivendi before they can form a people in ways that would lead them to affirm justice as fairness as the most appropriate conception of justice for them.[30]

§XI.4: "And very good it was"

I shall not use this last section to summarize the arguments of this book, or to reproduce in compressed form the picture of Rawls's view that emerges from my interpretation. Instead, I want to return to an important theme that I have touched on at critical junctures.

In §X.8, I said that Rawls's later account of stability—like his earlier one—stands in sharp contrast to those of Hobbes and of the two thinkers Rawls refers to as the "dark minds of Western thought," Augustine and Dostoevsky.[31]

27. See Rawls, *Lectures on the History of Moral Philosophy*, pp. 305–6; also §5.4.

28. Samuel Scheffler, *Boundaries and Allegiances* (Oxford: Oxford University Press, 2001), p. 147.

29. Samuel Freeman, "Original Meaning, Democratic Interpretation, and the Constitution," *Philosophy and Public Affairs* 21 (1992): pp. 3–42, pp. 11ff.

30. Cf. Thomas Nagel, "The Problem of Global Justice," *Philosophy and Public Affairs* 33 (2005): pp. 113–47, p. 147.

31. Rawls, *Lectures on the History of Political Philosophy*, p. 302.

In *PL* as in *TJ*, Rawls argues that justice as fairness would be stable by arguing that just institutions develop our natural capacities and shape our conceptions of what is good, so that we acquire, affirm, and preserve a sense of justice. His later account of stability, like his earlier one, is therefore intended to show that a just society suits our nature. Thus Rawls thought that if *PL*'s treatment of stability is plausible, or if *TJ*'s had been, we should take away from it an encouraging conception of what human beings are and can be. This brings to light one of the reasons he took up the question of stability in the first place, a reason the force of which emerges in seeing how Rawls would respond to a pointed criticism of his approach to political philosophy.

The aim of political philosophy, as Rawls conceives it, is practical. The task he ostensibly took up in *TJ* was that of formulating a conception that can serve as an enduring "foundation charter" for a well-ordered liberal democracy (*TJ*, p. 11/10). That Rawls set himself this task suggests that he thought the task of political philosophy is *immediately* practical: that it is of relevance to problems of distributive justice in contemporary liberal democracies. This suggestion is, of course, correct and it shows one of Rawls's reasons for taking up questions of stability. If a conception of justice did not generate its own support, then it would either fail to well-order society for long or it would have to be maintained by deceptive or repressive measures that would themselves be unacceptable to parties in the OP. In either case, the conception would prove unworkable as a liberal conception of justice.

But Rawls thinks that political philosophy is practical in other and deeper ways as well, and these also raise the question of whether justice as fairness would be inherently stable or stable for the right reasons. Some of the deeper practical tasks of political philosophy are discussed in his *Lectures on the History of Political Philosophy*,[32] but another—hinted at in *PL*—is best teased out of his published lectures on Kant.

Speaking of Kant in his *Lectures on the History of Moral Philosophy*, Rawls says

> he believes that we cannot sustain our devotion to the moral law, or commit ourselves to the advancement of its a priori object, the realm of ends or the highest good as the case may be, unless we firmly believe that its object is possible.

I think what Rawls has in mind is something like this. A "devotion to the moral law" demands commitment to a pattern of conduct—indeed, to a way of life—that requires discipline and perseverance in the face of temptation. This commitment will be too difficult to sustain if we cannot see the point of that commitment. We will be unable to see its point unless we believe that the realm of ends or the highest good can be realized, and realized in part through our efforts.

32. Rawls, *Lectures on the History of Political Philosophy*, pp. 10f.

Belief in the possibility of the realm of ends or the highest good is an example of "practical faith". But these possibilities cannot be the only objects of that faith. There are other things we must believe – there must be other articles of our faith—if we are to sustain our faith in these possibilities. What else must we believe, what other articles of faith must we accept, to sustain our practical faith that it is possible to realize a realm of ends and our devotion to the moral law?

Rawls replies that "we can believe that a realm of ends is possible in the world only if the order of nature and social necessities are not unfriendly to that ideal."[33] That the "order of nature" includes *human* nature is clear from Rawls's remark that practical faith "require[s] certain beliefs about *our* nature and the social world."[34] So according to Rawls, Kant thinks that we can sustain our commitment to the moral law only if we believe that human nature is not unfriendly to the realization of a realm of ends in the world. Similarly, I believe, Rawls thinks that we—and I take the "we" to refer to both members of the WOS and Rawls's readers—can sustain our commitment to the principles of justice and to bringing about a just society only if we think human nature is not unfriendly to the realization of that society in the world.

The amenability of human nature to the realization of a WOS is, Rawls thinks, an article of practical faith. One of the practical tasks of philosophy is that of showing that it is reasonable to accept that article of faith. Philosophy can show that human nature is "not unfriendly" to the realization of the WOS by showing that, at least under reasonably favorable conditions of a just society, human nature is such that we can develop the sentiments needed to maintain it. The conclusion that justice as fairness would be inherently stable is, Rawls thinks, part of what needs to be shown to show the reasonableness of our practical faith and our commitment to justice. But why think it needs to be *shown* that our nature is not unfriendly to the realization of a just society?

History raises grave doubts about whether human nature is in fact hostile to justice; indeed, according to Rawls, Kant worried that history can "arouse loathing for our species."[35] Those doubts may be heightened by more horrific events that Kant did not live to see. Both *TJ* and *PL* are supposed to address the doubts about us that recent history raises. Rawls makes this clear in the "Introduction to the Paperback Edition" of *PL*, in passage I shall quote at length because I shall later draw attention to an interesting change Rawls made in it. Rawls says:

> The wars of [the 20th] century with their extreme violence and increasing destructiveness, culminating in the manic evil of the Holocaust, raise in an acute way the question of whether political

33. Rawls, *Lectures on the History of Moral Philosophy*, p. 319.

34. Rawls, *History of Moral Philosophy*, p. 319 (emphasis added).

35. Rawls, *History of Moral Philosophy*, p. 320; cf. Robert Nozick, *The Examined Life* (New York: Simon and Schuster, 1989), pp. 236–42.

> relations must be governed by power and coercion alone. If a reasonably just society that subordinates power to its aims is not possible and people are largely amoral, if not incurably cynical and self-centered, one might ask with Kant whether it is worthwhile for human beings to live on the earth? We must start with the assumption that a reasonably just society is possible, and for it to be possible, human beings must have a moral nature, not of course a perfect such nature, yet one that can understand, act on, and be sufficiently moved by a reasonable conception of right and justice to support a society guided by its ideals and principles. *TJ* and *PL* try to sketch what the more reasonable conceptions of justice for a democratic regime are and to present a candidate for the most reasonable. They also consider how citizens need to be conceived to construct those more reasonable conceptions, and what their moral psychology has to be to support a reasonably just political society over time. (*PL*, p. lxii)

We have seen that both *TJ* and *PL* try to show how it is possible for a "reasonably just political society" to remain inherently stable—or stable for the right reasons—"over time." Since the possibility of a stably just society depends upon our having "a moral nature," the stability arguments of *TJ* and *PL* confirm that we do indeed have such a nature, and hence that we are not "largely amoral" or "incurably cynical or self-centered." Showing this, in turn, shows that political relations need not "be governed by power and coercion alone," nor stabilized in the ways that Augustine, Hobbes, and Dostoevsky thought they must be. Showing it therefore goes some way toward addressing the question Rawls says is raised by the Holocaust, by our two world wars and by our many lesser ones, and toward vindicating our practical faith.

Why think those questions need to be addressed *to sustain our commitment to justice*? To see the answer, consider some recent work by Raymond Geuss in which Geuss asks, in effect, whether Rawls's response to the violence and evil of the twentieth century was the right sort of response for philosophy to make. Geuss writes:

> What . . . would one have to believe about the world to think that "What is the correct conception of justice?" is the appropriate question to ask in the face of concentration camps, secret police, and the firebombing of cities? Are reflections about the correct distribution of goods and services in a "well-ordered society" the right *kind* of intellectual response to slavery, torture, and mass murder?[36]

Contra Geuss, reflections about our ability to sustain a just society are precisely the right kind of intellectual response if the only alternative is the response Geuss himself makes, which is to ask whether "political philosophy

36. Raymond Geuss, *Outside Ethics* (Princeton, NJ: Princeton University Press, 2005), p. 31.

[should] really be essentially about questions of fairness of distribution of resources" and whether "security and the control of violence [aren't] far more important." Geuss does not make clear what standards of importance are to be used in addressing his question. It is hard to resist the thought that Geuss believes questions about distributive justice are less important than questions about the control of violence because he thinks history shows human nature to be such that a commitment to justice is either utopian, or a luxury we cannot afford.

It is just *this* "kind of intellectual response to slavery, torture, and mass murder"—one that weakens our moral commitments in the name of political realism—that Rawls thinks has to be resisted. Resisting it is thus another one of the practical tasks Rawls assigns to political philosophy. Rawls tries to accomplish that task by showing that the realist conception of our nature—in twentieth-century forms of realism, a conception of our nature that is deeply indebted to Hobbes and Augustine[37]—is mistaken. We have seen that Rawls tried to show, against this conception, that human nature is good enough to "understand, act on, and be sufficiently moved by a reasonable conception of right and justice to support a society guided by its ideals and principles." Believing that we are good enough to do that, in the face of historical evidence to the contrary, is necessary to sustain a commitment to building a more just political world in the face of temptation, not only to injustice, but also to cynicism and despair.

The argument that human nature is good, or that it is at least good enough "to support a reasonably just political society over time," does not just affect our view of the political world and its possibilities.[38] "The answer we give to the question of whether a just democratic society is possible and can be stable for the right reasons," Rawls says "affects our background thoughts and

37. The classical source of twentieth-century Christian realism is, of course, Reinhold Niebuhr; for the influence of Augustine, see his "Augustine's Political Realism" in *The Essential Reinhold Niebuhr* (New Haven, CT: Yale University Press, 1986), ed. Robert McAfee Brown, pp. 123–41. For an accessible treatment, see Arthur M. Schlesinger, "Forgetting Reinhold Niebuhr," *The New York Times Book Review*, September 18, 2005.

38. A remark Rawls makes about Rousseau is also suggestive. He says,

> Rousseau's belief that human nature is good, and that it is through institutions that we become bad, comes to these two propositions:
> (a) Social institutions and the conditions of social life exercise a predominant influence over which human propensities will develop and express themselves over time. When realized, some of these propensities are good and some bad.
> (b) There exists at least one possible and reasonably workable scheme of legitimate political institutions that both satisfies the principles of political right and meets the requirements for institutional stability and human happiness.

The remark occurs at *Lectures in the History of Political Philosophy*, p. 206. Rawls does not say he agrees with Rousseau's belief that human nature is good. But if we suppose that he does, then we can read his work as, among other things, a sustained attempt to argue for the goodness of humanity.

attitudes about the world as a whole" (*PL*, p. lxi). Recalling that human nature is part of the "order of nature" helps us see why this is so.

According to the *Book of Genesis*, God surveyed each day's work "and saw that it was good." After the work of creation was complete, "God saw everything that he had made and, behold, it was very good"—or, as Rawls renders the passage in his undergraduate thesis "and so it was God saw all that He had made, and very good it was."[39] The judgment that "very good it was" clearly expresses God's "attitudes toward the world as a whole," but the judgment itself is very perplexing. It is not at all clear what it might mean to say that the world as a whole is "very good," for it is not clear what kind of value could characterize the world or what the truth-conditions of this judgment are. Even if we can ascertain the truth-conditions, we may not be in a position to see that they obtain, since God's point of view on creation is not one we can adopt. Still, if Rawls is right, we too can have "attitudes toward the world as a whole."

Augustine puzzled over how God could judge the world to very good, knowing that human beings would fall.[40] John Calvin thought that since God judged the world to be very good, we are bound "to acquiesce without controversy."[41] Even if we concur with God's judgment, however, we may differ with Calvin because we prefer to concur with that judgment on grounds of our own. The thought that the world as a whole is very good may strike us when we contemplate the grandeur and beauty, or the order, of the world we know. But in what way is the world very good? And what must we believe about the world to regard it as very good in that way?

In one of his lectures on Kant, Rawls remarks "What gives a view a religious aspect, I think, is that it has a conception of the world as a whole that presents it as in certain respects holy, or else as worthy of devotion and reverence."[42] I shall assume that holiness, or worthiness of reverence, are the kinds of value Rawls thinks we can take the world as a whole to have. But it is hard

39. John Rawls, *A Brief Inquiry into the Meaning of Sin and Faith* (Cambridge, MA: Harvard University Press, 2009), ed. Thomas Nagel, p. 137. The biblical passage is from Genesis 1:31, which I have quoted from the King James Version. Rawls's rendering does not match that of any of the English-language translations of the Bible that I have consulted, nor would it be an accurate English translation of the verse as found in the Vulgate or in Luther's German version. My guess is that Rawls quoted the familiar passage from memory, inadvertently altering it slightly so that the cadences of the verse fit his own sense of poetry.

40. See Augustine's commentary on the passage in *de Genesi ad Litteram*.

41. In the original Latin, Calvin's comment reads "Neque enim quod probavit, fas nobis est disputare, probarine debeat: sed potius absque controversia subscribere convenit." I am grateful to Richard Weaver for supplying me with this passage in correspondence and to Randall Zachmann for helpful correspondence about it. The standard translation of the passage, by John King, can be found online at http://en.wikisource.org/wiki/Calvin%E2%80%99s_Commentaries%E2%80%94Complete/Genesis_1–23/Chapter_1

42. Rawls, *Lectures on the History of Moral Philosophy*, p. 160.

to know what it means to say that we regard the world as having these kinds of value. Perhaps we will think it is an attitude that resists any further analysis or expression. Even if we do, we can still identify some of the necessary conditions of that attitude. Since the order of nature includes human nature, we can regard the world as a whole as holy, or as worthy of devotion and reverence, only if we recognize that our nature is part of what makes the world of nature that way. We can recognize that only if we regard ourselves as capable of living rightly. The task of showing that we have a moral nature is therefore necessary to show how we, too, can judge the world as a whole to be very good.

That Rawls saw the task of *TJ* and *PL* in precisely these terms is suggested by an unpublished version of the introduction to the paperback edition of *PL*, where the lengthy passage I quoted earlier in this section differs tellingly from the published version. Both versions of the paragraph begin with the same two sentences.

> The wars of [the 20th] century with their extreme violence and increasing destructiveness, culminating in the manic evil of the Holocaust, raise in an acute way the question of whether political relations must be governed by power and coercion alone. If a reasonably just society that subordinates power to its aims is not possible and people are largely amoral, if not incurably cynical and self-centered, one might ask with Kant whether it is worthwhile for human beings to live on the earth?

The published version continues "We must start with the assumption that a reasonably just society is possible, and for it to be possible, human beings must have a moral nature[.]" In the unpublished version, however, Rawls writes "*These thoughts quickly lead to a question not unrelated to the question of theodicy. It is said that after fashioning the world God saw that it was good. (Genesis 1) If it is good, a reasonably just society must be possible; and for it to be possible,* human beings must have a moral nature."[43] Rawls then implies, as he does in the published version, that *TJ* and *PL* try to vindicate faith in our having such a nature—against historical evidence to the contrary—by showing how a just society is possible.

But the lecture on Kant that I quoted a moment ago suggests that showing we *have* a moral nature is not enough for us to judge that a world which includes human beings is a good world. Instead, Rawls reads Kant as saying that our lives must *express* our moral nature. He says "our life in the world, and the world itself, lose their meaning and point" unless we "follow[] the moral law as it applies to us," "striv[e] to fashion in ourselves a firm good will," and "shap[e] our social world accordingly."[44] *TJ* and *PL* try to show that in a just

43. Rawls hand-dated my copy of this draft September 25, 1995; the italics are mine. David Reidy tells me that my correspondence with Rawls can be found in Box 1, Accession 15085, of the Rawls archive at Pusey Library, Harvard University.

44. Rawls, *Lectures on the History of Moral Philosophy*, pp. 160–1.

society, we would express our nature by conducting ourselves as free and equal persons or as free and equal citizens. The arguments for stability in both works turn on the conclusion that in a WOS, we would be the kind of persons who value the expression of our nature and for that reason "striv[e] to fashion in ourselves a *firm* good will" by preserving our sense of justice. *TJ* and *PL* therefore try to show that it is reasonable to believe what we must believe, and do what we must do, if we are to regard a world that includes us as a very good world.

I have argued that Rawls became dissatisfied with *TJ*'s treatment of stability because *TJ* failed to show that members of the WOS would all judge it good to preserve their "firm good will." His desire to correct that failure, and to show that they would maintain their sense of justice, provides one answer to the title question of this book, "Why political liberalism?". The appeal of the view that resulted, and the defense that can be offered for it, provide another. But if the reading of Rawls's view that I have offered in this book is right, then the arguments of this section suggest yet another and more surprising answer to that question. Political liberalism as Rawls develops it can help us to understand and affirm the very puzzling judgment that God is said to have passed upon the world.

This page intentionally left blank

Bibliography

Abbey, Ruth. *The Return of Feminist Liberalism*. Manuscript in progress.

Ackerman, Bruce. "Political Liberalisms." *The Journal of Philosophy* 91 (1994): 364–86.

Audard, Catherine. *John Rawls*. Montreal: McGill-Queens University Press, 2007.

Augustine. *Confessions*, trans. R. S. Pine-Coffin. New York: Penguin Books, 1961.

———. *de Civitate Dei*, trans. Henry Bettenson. New York: Penguin Books, 1972.

———. *de Genesi ad Litteram Libri Duodecim*. Online. Available: http://www.sant-agostino.it/latino/genesi_lettera/index.htm

Axelrod, Robert. *The Evolution of Cooperation*. New York: Basic Books, 1984.

Baird, Douglas G., Robert H. Gertner, and Randall C. Picker. *Game Theory and the Law*. Cambridge, MA: Harvard University Press, 1994.

Barry, Brian. "John Rawls and the Search for Stability." *Ethics* 105 (1995): 874–915.

Berlin, Isaiah. *The Crooked Timber of Humanity*. ed. Henry Hardy. New York: Alfred A. Knopf, 1991.

Bird, Colin. Review of John Rawls, *Lectures on the History of Political Philosophy*. *Ethics* 117 (2007): 784–90.

Bou-Habib, Paul. Review of *The Cambridge Companion to Rawls.* ed. Samuel Freeman. Cambridge: Cambridge University Press, 2004. *Journal of Moral Philosophy* 1, 3 (2004): 375–79.

Calebresi, Guido and A Douglas Melamed. "Property Rules, Liability Rules and Inalienability: One View of the Cathedral." *Harvard Law Review* 85, 6 (1972): 1089–1128.

Calvin, John. *Calvin's Commentaries—Complete*. Available: http://en.wikisource.org/wiki/Calvin%27s_Commentaries%E2%80%94Complete

Cohen, Gerald. *Karl Marx's Theory of History: A Defence*. Princeton, NJ: Princeton University Press, 1978.

———. "Where the Action Is: On the Site of Distributive Justice." *Philosophy and Public Affairs* 26 (1997): 3–30.

Cohen, Joshua. "Democratic Equality." *Ethics* 99 (1989): 727–51.
———. "Taking People as They Are." *Philosophy and Public Affairs* 30 (2001): 363–86.
———. "For a Democratic Society," in *Cambridge Companion to Rawls*, ed. Samuel Freeman, pp. 86–138.
Daniels, Norman, ed. *Reading Rawls*. Oxford: Basil Blackwell, 1975.
Darwall, Stephen. "A Defense of the Kantian Interpretation." *Ethics* 86 (1976): 164–70.
Davidson, Arnold. "Is Rawls a Kantian?" *Pacific Philosophical Quarterly* 66 (1985): 48–77.
Dreben, Burton. "On Rawls and Political Liberalism," in the *Cambridge Companion to Rawls*, ed. Samuel Freeman, pp. 316–46.
Dworkin, Ronald. "The Original Position," in *Reading Rawls*, ed. Norman Daniels. Oxford: Basil Blackwell, 1975, pp. 16–52.
Estlund, David. "The Survival of Egalitarian Justice in John Rawls's *Political Liberalism*." *Journal of Political Philosophy* 4 (1996): 68–78.
Forst, Rainer. *Contexts of Justice: Political Philosophy Beyond Liberalism and Communitarianism*. Berkeley, CA: University of California Press, 1994.
Freierson, Patrick. *Freedom and Anthropology in Kant's Moral Philosophy*. Cambridge: Cambridge University Press, 2003.
Freeman, Samuel. "Reason and Agreement in Social Contract Views." *Philosophy and Public Affairs* 19 (1990): 122–57.
———. "Original Meaning, Democratic Interpretation, and the Constitution." *Philosophy and Public Affairs* 21 (1992): 3–42.
———, ed. *Cambridge Companion to Rawls*. Cambridge: Cambridge University Press, 2003.
———. *Justice and the Social Contract.* New York: Oxford University Press, 2007.
Galston, William. "Moral Personality and Liberal Theory: John Rawls's 'Dewey Lectures'." *Political Theory* 10 (1982): 492–519.
Gauthier, David. *Morals by Agreement.* New York: Oxford University Press, 1986.
Geuss, Raymond. *Outside Ethics*. Princeton, NJ: Princeton University Press, 2005.
Gibbons, Robert. *Game Theory for Applied Economists.* Princeton, NJ: Princeton University Press, 1992.
Gray, John. *Two Faces of Liberalism*. New York: The New Press, 2000.
Gutmann, Amy. "The Communitarian Critique of Liberalism." *Philosophy and Public Affairs* 14 (1985): 308–22.
Hampton, Jean. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press, 1986.
Hauerwas, Stanley and William Willamon. *Resident Aliens: Life in the Christian Colony*. Nashville, TN: Abingdon Press, 1989.
Hittinger, Russell. "John Rawls: *Political Liberalism*." *The Review of Metaphysics* 47, 3 (1994): 585–602.
Hobbes, Thomas. *Leviathan*. 1651. Reprint eds. Richard E. Flatham and David Johnston. New York: W.W. Norton, 1997.
Holmes, Stephen. "The Gate Keeper." *The New Republic*, October 11, 1993, pp. 39–47.
Jones, L. Gregory. "Should Christians Affirm Rawls's Justice as Fairness? A Response to Professor Beckley." *Journal of Religious Ethics* 16 (1988): 251–71.
Kavka, Gregory S. Review of David Gauthier, *Morals by Agreement*. Oxford: Oxford University Press, 1986. *Mind* 96 (1987): 117–21.
Klosko, George. "Rawls's 'Political Philosophy' and American Democracy." *The American Political Science Review* 87 (1993): 348–59.

Korsgaard, Christine. "The Right to Lie: Kant on Dealing with Evil." *Philosophy and Public Affairs* 15 (1986): 325–349.
———. "Personal Identity and the Unity of Agency: A Kantian Response to Parfit." *Philosophy and Public Affairs* 18 (1989): 101–132.
———. *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996.
———. *Locke Lectures*. Online. Available: http://www.people.fas.harvard.edu/~korsgaar/
Krasnoff, Larry. "Consensus, Stability and Normativity in Rawls's *Political Liberalism*." *The Journal of Philosophy* 95, 6 (1998): 269–92.
Larmore, Charles. "Political Liberalism." *Political Theory* 18 (1990): pp. 339–60.
———. "Public Reason," in *Cambridge Companion to Rawls*, ed. Freeman, pp. 368–93.
———. "The Moral Basis of Political Liberalism," in Charles Larmore, *The Autonomy of Morality* (Cambridge University Press, 2008).
———. *The Autonomy of Morality.* Cambridge: Cambridge University Press, 2008.
McClennan, Edward. "Justice and the Problem of Stability." *Philosophy and Public Affairs* 18 (1989): 3–30.
McMullin, Ernan. "Values in Science." *PSA 1982*, 2: pp. 3–28.
Moore, Margaret. *Foundations of Liberalism*. New York: Oxford University Press, 1993.
Nagel, Thomas. "Moral Conflict and Political Legitimacy." *Philosophy and Public Affairs* 16 (1987): 215–40.
———. "The Problem of Global Justice." *Philosophy and Public Affairs* 33 (2005): 113–47.
Nelson, Alan. "Economic Rationality and Morality." *Philosophy and Public Affairs* 17 (1988): 149–66.
Newman, John Henry. *An Essay in the Development of Christian Doctrine*. Whitefish, MT: Kessinger Publishing, 2007.
Niebuhr, Reinhold. *The Essential Reinhold Niebuhr*, ed. Robert McAfee Brown. New Haven, CT: Yale University Press, 1986.
Noonan, John T., Jr. *The Luster of Our Country: The American Experience of Religious Freedom*. Berkeley, CA: University of California Press, 2000.
Nozick, Robert. *The Examined Life.* New York: Simon and Schuster, 1989.
Nussbaum, Martha. *Women and Human Development*. Cambridge: Cambridge University Press.
Parfit, Derek. "Later Selves and Moral Principles," in *Philosophy and Personal Relations: An Anglo-French Study*, ed. Alan Montefiore. London: Routledge & Kegan Paul, 1973, pp. 137–169.
Pogge, Thomas. "The Kantian Interpretation of Justice as Fairness." *Zeitschrift für philosophische Forschung* 35 (1981): 47–65.
Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971, 1999.
———. *Political Liberalism.* New York: Columbia University Press, 1996.
———. *Collected Papers*, ed. Samuel Freeman. Cambridge, MA: Harvard University Press, 1999.
———. *Lectures on the History of Moral Philosophy*, ed. Barbara Herman. Cambridge, MA: Harvard University Press, 2000.
———. *Law of Peoples*. Cambridge, MA: Harvard University Press, 2001.
———. *Justice as Fairness: A Restatement*, ed. Erin Kelly. Cambridge, MA: Harvard University Press, 2001.
———. *Lectures on the History of Political Philosophy*, ed. Samuel Freeman. Cambridge, MA: Harvard University Press, 2007.

———. *A Brief Inquiry into the Meaning of Sin and Faith*, ed. Thomas Nagel. Cambridge, MA: Harvard University Press, 2009.
Richardson, Henry. "John Rawls." *Internet Encyclopedia of Philosophy*. Online. Available: http://www.iep.utm.edu/rawls/
Rorty, Richard. "Of Trotsky and the Wild Orchids," in his *Philosophy and Social Hope*. New York: Penguin Books, 1999, pp. 3–20.
Sandel, Michael. *Liberalism and the Limits of Justice.* Cambridge: Cambridge University Press, 1982.
Scheffler, Samuel. *Boundaries and Allegiances*. Oxford: Oxford University Press, 2001.
Schlesinger, Arthur M. "Forgetting Reinhold Niebuhr." *The New York Times Book Review*, September 18, 2005. Online. Available: http://www.nytimes.com/2005/09/18/books/review/18schlesinger.html
Sen, Amartya K. "Isolation, Assurance and the Social Rate of Discount." *The Quarterly Journal of Economics* 81, 1 (1967): 112–24.
———. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs*, 6 (1977): 317–44.
———. "Why Exactly is Commitment Important for Rationality?" *Economics and Philosophy* 21 (2005): 5–13.
Skyrms, Bryan. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press, 1993.
Steele, Gerry R. "Understanding Economic Man: Psychology, Rationality and Values." *American Journal of Economics and Sociology* 63 (2004): 1021–55.
Ullmann-Margalit, Edna. *The Emergence of Norms.* Oxford: Oxford University Press, 1977.
Waldron, Jeremy. "Disagreements about Justice." *Pacific Philosophical Quarterly* 75 (1994): 372–87.
Weithman, Paul. "Augustine and Aquinas on Original Sin and the Purposes of Political Authority." *Journal of the History of Philosophy* xxx (1992): 353–76.
———. "Augustine's Political Thought," in *Cambridge Companion to Augustine*, eds. Eleonore Stump and Norman Kretzmann. Cambridge: Cambridge University Press, 2001, pp. 234–52.
———. Review of Jeffrey Stout. *Democracy and Tradition*. (Princeton, NJ: Princeton University Press, 2004). *Faith and Philosophy* 23 (2006): 221–29.
———. "John Rawls and the Task of Political Philosophy." *The Review of Politics* 71 (2009): 113–125.
———. Review of John Rawls. *A Brief Inquiry into the Meaning of Sin and Faith*, ed. Thomas Nagel. Cambridge, MA: Harvard University Press, 2009. *Notre Dame Philosophical Reviews*. August 18, 2009. Online. Available: http://ndpr.nd.edu/review.cfm?id=17045
———. "Claims and Capabilities." *The Philosophy of Martha Nussbaum*, ed. Randall Auxier. Chicago, IL: Open Court Publishing, forthcoming.
Wenar, Leif. "The Unity of Rawls's Thought." *Journal of Moral Philosophy* 1 (2004): 265–75.
Williams, Bernard. *In the Beginning Was the Deed*. Princeton, NJ: Princeton University Press, 2005.
Wittgenstein, Ludwig. *Tractatus Logico-Philosophicus*, trans. D. F. Pears and B. F. McGuinness. London: Routledge and Kegan Paul, 1974.
Wolin, Sheldon. Review of John Rawls. *Political Liberalism*. *Political Theory* 24 (1996): 97–129.

Index