Justice, Constructivism, and the Egalitarian Ethos

Explorations in Rawlsian Political Philosophy

A. Faik Kurtulmus

The Queen's College

Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Politics

Department of Politics and International Relations

University of Oxford

Hilary 2010

Approx. 78,000 Words

For Çiçek and Merve

Abstract

This thesis defends John Rawls's constructivist theory of justice against three distinct challenges.

Part one addresses G. A. Cohen's claim that Rawls's constructivism is committed to a mistaken thesis about the relationship between facts and principles. It argues that Rawls's constructivist procedure embodies substantial moral commitments, and offers an intra-normative reduction rather than a metaethical account. Rawls's claims about the role of facts in moral theorizing in *A Theory of Justice* should be interpreted as suggesting that some of our moral beliefs, which we are inclined to hold without reference to facts, are, in fact, true, because certain facts obtain. This thesis and the acknowledgement of the moral assumptions of Rawls's constructivism help to show that Rawls does not, and does not need to, deny Cohen's thesis.

Part two defends the characterization of the decision problem in Rawls's original position as a decision problem under uncertainty. Rawls stipulates that the denizens of the original position lack information that they could use to arrive at estimates of the likelihood of ending up in any given social position. It has been argued that Rawls does not have good grounds for this stipulation. I argue that given the nature of the value function we should attribute to the denizens of the original position and our cognitive limitations, which also apply to the denizens of the original position, their decision problem can be characterized as one under uncertainty even if we stipulate that they know that they have an equal chance of being in any individual's place.

Part three assesses the claim that a true commitment to Rawls's difference principle requires a further commitment to an egalitarian ethos. This egalitarian ethos is offered as a means to bring about equality and Pareto-optimality. Accordingly, I try to undermine the case for an egalitarian ethos by challenging the desirability of the ends it is supposed to further or by showing that it is redundant. I argue that if primary goods are the metric of justice, then Pareto optimality in the space of the metric of justice is undesirable. I then argue that if the metric of justice is welfare, depending on the theory of welfare we adopt, an egalitarian ethos will either be redundant or will have objectionably paternalistic consequences.

Acknowledgements

My supervisor David Miller's conscientious supervision, incisive criticism, and patience has made this thesis much better than it would otherwise have been.

Conversations with fellow students and their comments on parts of this thesis have taught me a lot about how to do philosophy and its rewards. I wish to thank in particular Rob Jubb, Seth Lazar, Ben Saunders, and Nicholas Vrousalis. Non-philosopher friends have also been subjected to parts of my thesis, and my incessant talking about philosophical topics. Among them, I owe special thanks to Levent Akkok, Oguz Erdin, Evren Lus, Hilmi Lus, Olgun Ural, and Hikmet Unlu.

I wish to thank David Christensen and Brad Hooker who have kindly read and commented on parts of my thesis dealing with their work even though we have never met in person.

Much of this thesis deals with the ideas of Jerry Cohen. I knew what a great philosopher he was before coming to Oxford. Meeting him I found out what a great human being he was. He was always generous, kind, fun to be around, and single minded in his pursuit of the truth. His comments on the parts of my thesis he read have made me rethink and reformulate many of my arguments, and altogether drop many others. I will miss him.

Andrew Williams and Adam Swift examined this thesis, and provided extensive and extremely valuable comments.

I wish to express my gratitude for the financial support of the Clarendon Fund and the Department of Politics and International Relations which made this thesis possible.

I am deeply grateful to my parents for the financial and emotional support they have provided. I'm also grateful to my mother-in-law whose assistance in the past two years since Çiçek was born enabled me to find the time to focus on my thesis. Finally my greatest debt is to my wife, Merve. None of this would have been possible without her support. This thesis is dedicated, with love, to her and our daughter Çiçek.

Contents

Chapter 1

Introduction

1.1 Introduction

This thesis is not the one I intended to write when I arrived at Oxford. Before arriving here, I had written an MA thesis on Rawls's *Law of Peoples* where I had argued that Rawls's theory there followed from the approach to justification that Rawls had employed in *Political Liberalism*, and that Rawls's main concern in both books was legitimacy. For someone like me, coming from an illiberal society, both books were disappointments. I was, as David Miller aptly puts it, 'yearning for Rawlsian "moral geometry",' and, oddly enough, not finding enough of it in Rawls's later works.[1] It's all fine and well to have arguments for the people of liberal countries, I thought, but what about people in illiberal countries? I thought there had to be arguments that I could provide to my illiberal compatriots that would conclusively establish the rightness of egalitarian liberalism. I wanted arguments

[1] David Miller, *Social Justice*, (Oxford: Clarendon Press, 1976), p. 343.

that would *force* them to that conclusion.[2] So my plan was to go back to *A Theory of Justice* and to re-establish this 'moral geometry'. Of course, this is not what happened. Instead, I came across G. A. Cohen's critique of Rawls's constructivism and theory of justice. This critique came with 'the added persuasiveness of personal presentation', which in Jerry's case was remarkable.[3] Given these challenges, my initial project seemed hopeless. Instead, I set about addressing the two of his challenges that worried me the most: his claim that Rawls's constructivism rests on a mistaken thesis about the relationship between facts and principles, and his argument that the commitment to Rawls's difference principle requires commitment to an egalitarian ethos. My attempt to respond to these charges takes up parts 1 and 3 of this thesis. The only thing that remains of my initial plan is my defence of Rawls against the criticism, voiced by Hare, Nagel, and Parfit, that Rawls's characterization of the decision problem in the original position as one under uncertainty is unjustified. I respond to this challenge in part 2.

My grandiose ambitions were cut down to size. Nevertheless, I still do believe that despite all of the strong criticisms of Rawls's constructivism, its central idea is ultimately correct, and we can arrive at satisfactory responses to these criticisms, though the theory we end up with will fall short of a 'moral geometry'. My rather modest aim here is to make some headway in this project of vindicating Rawls's theory by responding to three influential criticisms of the theory.

[2]At the time I was too naïve to realize that 'the penalty philosophers wield is... rather weak'. Robert Nozick, *Philosophical Explanations*, (Oxford: Clarendon Press, 1981), p. 4.

[3]G. A Cohen, *If You're an Egalitarian, How Come You're so Rich?* (Cambridge, Mass: Harvard University Press, 2001), p. 18.

1.2 The Scope of My Defence

What I mean by Rawls's constructivism has a wider scope than it's often understood, so I should clarify what I mean by it. The central idea I wish to defend is that the correct, or the most reasonable, principles of justice are 'the principles that would be agreed to in an appropriate initial situation that is fair between individuals conceived as free and equal', and that Rawls's characterization of the veil of ignorance, save a few details, correctly identifies this initial condition of fairness.[4] One might want to ask *for whom* these principles are correct, or most reasonable. Is it for all people, or the citizens of democratic societies? I shall remain neutral on the answer to this question. The challenges I address apply with equal force whichever way we answer this question, and my response to these challenges will not assume an answer to these questions. One might also think that this procedure identifies not only the correct principles of justice, but the correct principles for the whole of morality. I shall remain neutral on whether this claim can be made good. Suffice it to note that the first two challenges I address apply to this extended application of the original position just as strongly.

Rawls, of course, has an elaborate framework that he employs in addition to the device of the original position. Though I do not wish to deny the value of the rest of Rawls's theory, my main focus is the use of the original position to arrive at the principles of justice. Only in the last part will I turn to substantive questions, and when I do address them, I try to rely as little

[4]John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 226.

as possible on the Rawlsian framework beyond what's strictly part of the original position device. For instance, when evaluating Cohen's argument for an egalitarian ethos, I do not assume that primary goods are the correct metric of justice, and my defence does not rely on the primacy of the basic structure as the subject of justice since neither of these views are a necessary consequence of the adoption of the veil of ignorance.

1.3 On the Appeal of Rawls's Constructivism

In *On the Plurality of Worlds*, David Lewis argues for his modal realism by arguing that the costs one incurs by accepting the existence of possible worlds is outweighed by the 'theoretical benefits' gained by the acceptance of the theory, and that no other theory outperforms his theory on a cost-benefit analysis.[5] This seems like the correct approach to adopt in evaluating a complex and wide-reaching theory like Rawls's. Since in this thesis I'm on the defensive, I shall not do much to establish the benefits of the theory. What I shall do, instead, is to lower the costs of accepting it. I seek to demonstrate that many of the costly commitments of Rawls's theory can be avoided, or that the attribution of costs are based on misunderstandings of the theory.

Before outlining some of the costs, which I will argue we can avoid, let me very briefly say something about the appeal of Rawls's constructivism. Its appeal, I think, rests on three interrelated ideas. The first idea Rawls's

[5]David K. Lewis, *On the Plurality of Worlds*, (Oxford: Blackwell, 2001)

constructivism draws its appeal from is the widely shared bedrock conviction that people are due equal concern and respect. What this fundamental commitment amounts to is difficult to establish, but the original position provides an intuitively plausible way to give content to this commitment. A second idea that contributes to the appeal of Rawls's constructivism, and flows from the first idea, is a commitment as Waldron puts it 'to a conception of freedom and of respect for the capacities and the agency of individual men and women' that generates 'a requirement that all aspects of the social world should either be made acceptable or be capable of being made acceptable to every last individual'.[6] Because Rawls's constructivism relies on a slender basis, and this basis is the one that gives rise to the requirement Waldron states, it has a good chance of satisfying this requirement. A third related idea is our conception of ourselves and others as free and equal. We take ourselves to have certain reasons in virtue of our self-conception.[7] By drawing a connection between our reasons we have in virtue of the fact that we hold ourselves and fellow human beings to be free and equal and the principles of justice, Rawls's constructivism illuminates the reason-giving force of principles of justice. These three ideas recommend Rawls's constructivism, but do not establish a conclusive case in its favour. As I shall argue in Part 1, the principles that emerge out of the original position have to be acceptable to us upon due reflection if the constructivist project is to succeed.

[6]Jeremy Waldron, 'Theoretical Foundations of Liberalism', *The Philosophical Quarterly*, 37 April (1987):147, p. 128.

[7]For an insightful discussion of how our conception of ourselves gives us reasons see Christine M Korsgaard, *The Sources of Normativity*, (Cambridge: Cambridge University Press, 1996), p. 100-103. There's also Rawls's discussion of this idea. See John Rawls, *Political Liberalism*, New edition. (New York: Columbia University Press, 1996), p. 84.

1.4 Lowering the Costs of Acceptance

What are the commitments of Rawls's theory that are actually, or argued to be, costly, and how do I propose to avoid them? One of the central claims of the first part of this thesis is that Rawls's constructivism is best evaluated within the general framework of moral and political theorizing that is given by Rawls's account of reflective equilibrium. One commitment that political philosophers don't worry much about, but is nevertheless one worth worrying about, is the characterization of reflective equilibrium as a coherentist account of epistemic justification. I argue that we can give up this characterization of reflective equilibrium. I suggest that reflective equilibrium is best understood as an account of the reasoning process of typical moral inquirers who already have moral beliefs at different degrees of generality, and of which they are confident to different degrees. By interpreting reflective equilibrium in this way, we are able to side-step debates about coherentism vs foundationalism and externalism about justification vs internalism about justification, and we can avoid the costs that come with these theories. In addition to helping us avoid these costs, this interpretation of reflective equilibrium is able to explain why it is so popular. As Harman notes, while reflective equilibrium is widely adopted in moral and political philosophy 'reflective equilibrium... [is] often ignored in contemporary epistemology, which is still focused on special foundationalism'.[8] If we don't interpret reflective equilibrium as an account of justification in competition with foundationalism, then this discrepancy

[8]Gilbert Harman, 'Three Trends in Moral and Political Philosophy', *The Journal of Value Inquiry*, 37 (2006):3. It worth noting that even though reflective equilibrium is not popular amongst epistemologists as an account of justification, the method is widely employed. My account of reflective equilibrium in chapter 2 can account for this fact.

can easily be explained.

Another costly commitment I suggest we can avoid is the supposed metaethical nature of Rawls's constructivism. There are, no doubt, parts of Rawls's text that suggests that he's offering a metaethical account, particularly in his Dewey Lectures 'Kantian Constructivism in Moral Theory'. However, as I argue in chapter 4, Rawls's constructivism is not a metaethical doctrine, and is compatible with various metaethical doctrines. It is best understood as offering an account of reasons in one domain in terms of reasons in another domain. In other words, Rawls is best understood as offering an intra-normative reduction. One is free to adopt any metaethical account of the reasons that functions as the reduction base. So, in adopting Rawls's constructivism we don't need to incur the costs associated with the adoption of a metaethical doctrine.

It is this mistaken characterization of Rawls's theory as a metaethical account–in addition to the reading Cohen offers of crucial passages in Rawls–that affords plausibility to Cohen's charge that Rawls denies his thesis that 'whenever a fact supports, that is, gives us reason to affirm, a principle, it does so in virtue of a more ultimate principle that is not supported by any facts'.[9] For if Rawls's constructivism is to go all the way down, and count as a *bona fide* metaethical account, then there shouldn't be moral principles that underlie Rawls's constructivist procedure. However, Rawls's

[9]G. A Cohen, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008), p. 20.

constructivism doesn't go all the way down.[10] Since Rawls is not offering a metaethical account, Rawls's constructivism can afford to embody normative commitments, and not to deny Cohen's thesis.

In chapters 3 and 4, I argue that the tasks of building a theory of justice that is best from our perspective, and building a theory that we can justify to others are distinct tasks. In our theorizing about the best theory from our perspective it can be epistemically rational for us to employ premises that others think are false. However, if we accept the liberal principle of legitimacy, which requires that the principles that govern the use of political power should be acceptable to the people over whom they are exercised, it would be unreasonable to employ these premises as grounds for principles that will be coercively imposed. Consequently, there are two distinct projects: the search for the theory of justice that is best from the individual perspective, and the search for a theory of principles which we can legitimately impose on others. These two projects are subject to different constraints and they are entitled to make use of different resources. Failing to observe this fact makes both projects appear more costly than they are. The project of determining the principles that pass the test of legitimacy can introduce considerations that are *prima facie* alien to the concept of justice if this helps it achieve its task. For instance, we might believe that welfare is the correct metric of justice, but concede that given the practical difficulties involved in putting this conception into practice entails, it makes more sense to rely on primary

[10]It has to be noted that Cohen seems to acknowledge this fact, when he talks of the principles that justify the original position. However, he, at the same time, takes the denial of his thesis to be essential to constructivism.Cohen, *Rescuing Justice*, pp. 231, 241.

goods. Even though the concern with feasibility might have no place in the theory of justice that best captures our convictions about justice, feasibility has a role in the project of arriving at principles that pass the test of liberal legitimacy.

Having *ad hoc* assumptions is another cost that a theory should avoid. In the case of Rawls's constructivism, the prominent *ad hoc* assumption is the claim that denizens of the original position lack probability estimates. The characterization of the decision problem in the original position as one of uncertainty is crucial for the derivation of Rawls's principles of justice, but Rawls seems to lack a good reason to defend this assumption. Since the design of the original position is up to Rawls, he can just as well stipulate that the denizens of the original position have an equal chance of ending up in any individual's position. This *ad hoc* move, I argue, can be avoided. In chapter 6 I suggest that in order to characterize the decision problem in the original position as one of uncertainty, we don't need to assume that the denizens of the original position lack probability estimates. Instead, we can offer reasons for why the value functions they employ contain some inherent uncertainty.

One final cost we can avoid is the cost of commitment to an egalitarian ethos. Suppose we hold that Rawls's derivation of his two principles of justice is correct, and the difference principle is a principle of justice. Cohen has argued that the commitment to the difference principle implies a further commitment to an egalitarian ethos that governs each individual's occupa-

tional choices. Rawlsians have wanted to resist Cohen's argument for an egalitarian ethos by arguing that Rawls's principles of justice apply only to the basic structure of society, or by arguing that principles of justice need to conform to certain publicity requirements that the egalitarian ethos cannot satisfy. Cohen has identified serious problems with this line of argument. Even though Rawlsians may ultimately respond to Cohen's challenges, it is profitable to find a line of argument that does not rely on the basic structure or the publicity restriction. In chapters 7 and 8, I offer a two-pronged argument against Cohen's egalitarian ethos that doesn't rely on the basic structure argument, or the publicity argument, and does not assume that primary goods are the correct metric of justice.

I don't seek to show that there's something inherently wrong about a theory of justice incorporating a moral duty that governs individuals' occupational choices such as which jobs to work in and how much to work. Instead, I examine Cohen's specific proposal by evaluating the ends it is meant to serve and its effectiveness in achieving those ends. Cohen's egalitarian ethos is offered as a means for the co-realization of equality and Pareto-optimality. Therefore, we can ask two questions to evaluate it. First, we can ask whether these two ends and their co-realization is desirable. Secondly, assuming that these two ends are desirable, we can ask whether an egalitarian ethos is able to bring about these ends, and whether it is really necessary to bring about these ends.

I argue that the co-realization of equality and Pareto-optimality when the metric of justice is primary goods is not morally desirable, because Pareto-optimality in the space of primary goods that are governed by the difference principle, namely income and wealth, is not desirable. Accordingly, an egalitarian ethos offered as a means to the realization of this goal is also not desirable. In chapter 8, I look at the case for an egalitarian ethos when welfare is assumed to be the metric of justice. This requires evaluating the case for an egalitarian ethos on different accounts of welfare. I argue that if the correct account of welfare is actual preference satisfaction, then an egalitarian ethos cannot bring about equality of welfare. If, however, we are in a position to bring about equality of welfare through other means, then the same means can bring about the co-realization of Pareto-optimality and equality. So, an egalitarian ethos is redundant. I then examine the case for an egalitarian ethos when either the objective list theory or the informed preference satisfaction theory gives the correct account of welfare. I suggest that depending on the methods we adopt in the pursuit of Pareto-optimality in the space of welfare so understood, either the redundancy point applies, or the egalitarian ethos has objectionably paternalistic consequences. Given the arguments of chapters 7 and 8, we can reject Cohen's argument for an egalitarian ethos without having recourse to theoretically costly commitments such as the publicity requirement, or the basic structure restriction.

1.5 A Comment About the Approach Taken Here

Two aspects of the approach I have taken in this thesis might strike the reader as unusual: the time I spend on questions about the nature of moral and political theorizing, and problems with interpersonal justification, and my reliance on literature in other fields, such as economics and philosophy of mind.

Why do I spend so much time trying to develop an account of reflective equilibrium and on questions of interpersonal justification? Part of the reason, I must concede, is autobiographical. I had to figure out for myself what it is reasonable to expect philosophical reflection and argumentation to establish. A second reason is that unless we have a clear account of moral theorizing and interpersonal justification, we will misunderstand the nature of Rawls's constructivism, and what it can accomplish. In particular, we will fail to recognize that the different projects of Rawls's constructivisms – that of arriving at the best theory of justice from the first person perspective and the project of arriving at principles that can be legitimately imposed – require different argumentative strategies, and are responsive to different constraints. In short, we will misunderstand the nature of Rawls's projects. A third reason is that implicit answers to these questions guide everyone's theorizing, but it is only if these implicit answers are made public that we can come to see their shortcomings. One cannot set out to state, let alone defend, all of one's assumptions. However, having our key assumptions set

out in public and open to direct challenges gives us a chance to improve on them.

Relying on findings from disciplines other than political philosophy has its benefits and its costs. While my arguments, if they are convincing, will be a testimony to these benefits, I should acknowledge the possible costs here. Narrow specialization enables one to acquire a command of the field one specializes in that would otherwise be impossible and contributes to the progress of the field. However, the practice of narrow specialization also has the consequence that the findings of fields where there's such specialization are hard to access and to assess for outsiders. Herein lies the cost of relying on fields other than one's speciality. How can one know that the findings from another field one relies on are well-established? Reading around the relevant literature, and looking for criticisms of the findings you rely on is a step. However, it's rare to find findings that gain universal assent, so there's an unavoidable risk. I believe the risk is one worth taking. We cannot expect the slender basis that we have specialized in to be always enough to sustain our arguments.

Here are the precautions I have taken in addition to looking at the criticisms of the findings and views I have relied on. Firstly, I have tried to rely on uncontroversial findings. When there were competing accounts, I have indicated this, and tried to formulate my account in ways that rely on claims shared by at least some of the competing positions. For instance, in chapter 6, I relied on the thesis that thinking takes place in a language. I have

taken care to make my argument compatible with both the claim that the language of thought is an innate language distinct from natural language, and the claim that the language of thought is natural language.[11] Secondly, I have taken steps to ensure that the literature I rely on is not denied by the proponents of the view I'm arguing against. For instance, in chapter 8, where I argue against egalitarian ethi, I relied on the literatures on incentives and optimal taxation. This literature is also relied on by the proponents of egalitarian ethos. In the end, whether the measures I have taken to limit the costs of relying on fields outside my specialty were enough, and whether the risk I've taken was one worth taking, is up to the reader to decide.

[11]The first position is represented by Jerry A. Fodor, *The Language of Thought*, (Cambridge, Mass: Harvard University Press, 1975). The latter position is represented by Peter Carruthers, *Language, Thought and Consciousness: An Essay In Philosophical Psychology*, (Cambridge: Cambridge University Press, 1996).

Part I

Reflective Equilibrium and Constructivism

Chapter 2

Reflective Equilibrium

.

> SOCRATES: Now by 'thinking' do you mean the same as I do?
> THEAETETUS: What do you mean by it?
> SOCRATES: A talk which the soul has with itself about the objects under its consideration...[T]his is the kind of picture I have of it. It seems to me that the soul when it thinks is simply carrying on a discussion in which it asks itself questions and answers them itself, affirms and denies. And when it arrives at something definite, either by a gradual process or a sudden leap, when it affirms one thing consistently and without divided counsel, we call this its judgement.
>
> Plato, *Theaetetus*, 189e-190a[1]

[1]M.J. Levett's translation revised by Myes Burnyeat in Plato; John M. Cooper and D. S Hutchinson, editors, *Plato: Complete Works*, (Indianapolis, IN: Hackett Publishing Company, 1997)

2.1 Introduction

It is often claimed that reflective equilibrium is a coherentist account of epistemic justification.[2] However, we have reasons to doubt this claim. Whereas reflective equilibrium is widely embraced by many philosophers in their practice, the same cannot be said of coherentist theories of epistemic justification. Coherentism has many critics.[3] Coherentism in epistemology tends to be allied with another thesis in epistemology, internalism. According to internalism about justification, very roughly, what justifies S's belief in p should be internal to S. This view is also contested in epistemology.[4] So, there's something puzzling. Why is reflective equilibrium so popular? There's also a cause for concern. If reflective equilibrium is interpreted as a coherentist account of epistemic justification, then the political philosopher who adopts it is adopting a risky philosophical commitment.

In this chapter I shall argue that we can interpret reflective equilibrium as an account of the reasoning process of a typical moral inquirer rather than a coherentist account of epistemic justification. This interpretation can easily account for the popularity reflective equilibrium enjoys, and help political philosophers carry on as usual. What I mean by 'the reasoning process of a

[2]The most influential expositor of reflective equilibrium, Norman Daniels, refers to reflective equilibrium as a coherentist account of moral justification in several places in his *Justice and Justification*. See Norman Daniels, *Justice and Justification: Reflective Equilibrium in Theory*, (Cambridge: Cambridge University Press, 1996), pp. 1, 3, 153.

[3]For a recent and powerful critique that argues against coherentism about justification by arguing that coherence alone is not truth conducive see Erik J Olsson, *Against Coherence: Truth, Probability, and Justification*, (Oxford: Oxford University Press, 2005).

[4]For a helpful collection of articles by the proponents of both internalism and externalism see Hilary Kornblith, *Epistemology: Internalism and Externalism*, (Malden, Mass: Blackwell Publishers, 2001).

typical moral inquirer' needs to be clarified. I take up this task in the next two sections. Section 2 offers an account of what I will call internal epistemic reasons, which are the epistemic reasons for belief available for an inquirer. In Section 3, I specify what I mean by a typical moral inquirer, and show how Rawls's account of reflective equilibrium is an illustration of the reasoning process of a typical moral inquirer. In Section 4, I demonstrate one way in which reflective equilibrium is philosophically edifying despite not being an account of epistemic justification. I argue that reflective equilibrium, by reminding us that our first-order and second-order beliefs about morality have the same standing, helps us see that metaethical theories which conflict with our strongly held first-order beliefs will not be acceptable to us. This last claim will bear on our discussions of constructivism in Chapter 4 where I will argue that Rawls's constructivism isn't a metaethical doctrine.

2.2 Internal Epistemic Reasons

In this section I provide an account of epistemic reasons for belief that a person has. This is the first step of our account of the reasoning process of typical moral inquirers, because reasoning is figuring out what one has reason to believe.

Let's say the door to my room is open. This fact is a reason for me to believe that 'The door to my room is open'. In this sense of reason, which I shall call external epistemic reasons for belief, every fact gives us a reason to

believe the proposition expressing it.[5] I call such reasons external because they need do not need to refer to a subject's mental states, unless they involve claims about that subject's mental states. The qualifier epistemic is intended to leave open the possibility that there may be non-epistemic reasons for believing a proposition. If the world will end if S believes that p, even though p is true S may have a non-epistemic reason to believe not-p. We can define external epistemic reasons for belief as follows.

> *External epistemic reasons for belief*: S has an external epistemic reason to believe p if and only if p is true.

External epistemic reasons capture one kind of statements about reason that we are inclined to make. However, they don't take account of the first-person perspective. There is another type of reason statements that are concerned precisely with the first-person perspective.

Suppose an evil demon has set things up in such a way that the door to my room appears closed, when in fact it is open. Furthermore, no matter how thoroughly and carefully I reason from premises and methods of reasoning and inquiry available to me, I can not arrive at the conclusion that the door to my room is open. In this sense of reason, I do not have a reason to believe that 'The door to my room is open'. Notice that in this sense of reason, which I shall call *internal epistemic reasons*, a reason for me to believe p is other beliefs of mine coupled with methods of reasoning available to me and

[5] As it should be clear, my discussion of reasons for belief is indebted to Williams's discussion of internal and external reasons. Bernard A. O. Williams, 'Internal and External Reasons', in: *Moral Luck: Philosophical Papers, 1973-1980*, (Cambridge: Cambridge University Press, 1981).

the epistemic norms I'm committed to. Let us call these beliefs and norms a person's *subjective belief and reasoning set*.

> *Internal epistemic reasons for belief*: S has an internal epistemic reason to believe p if and only if there is a sound deliberative route from S's subjective belief and reasoning set to the belief that p.

We need to fill in some of the details of the subjective belief and reasoning set (hereafter SBR) to arrive at a more accurate characterization of internal epistemic reasons for belief. A person's SBR contains their existing beliefs, their norms of reasoning, and their epistemic norms. Existing beliefs include all of the beliefs that an individual has. Included here are not only what we might call first-order beliefs but also beliefs about the faculties an individual has, the reliability of beliefs, what counts as evidence, which senses are reliable etc. I put the deliverances of sense perception also under beliefs. This may be problematic in certain contexts, but for our purposes this assumption is not problematic. Subjects have varying degrees of confidence in different beliefs.

Norms of reasoning include any rules of inference that take one from a set of premises to a conclusion. Norms of reasoning, here, include the norms of inference that the subject takes to be valid upon due reflection. They include things such as rules of deductive, inductive, and probabilistic reasoning, and inference to the best explanation. The rules of inference an individual follows need not be explicitly formulated by that person. In fact,

to formulate the rules of reasoning one follows may require extensive effort. What is required for a rule of inference to figure in an internal epistemic reason statement for an individual is that the rule *authoritatively* governs that individual's reasoning. The requirement is not that the rules actually do govern all the time, because that would rule out the possibility of a subject making errors in reasoning by her own lights. What is intended by a rule of inference governing authoritatively can be explained by drawing on a distinction from linguistics. Linguists distinguish between competence and performance. Speakers of a language know how to judge grammaticality, but may sometimes produce ungrammatical sentences. That is, they are competent judges of grammaticality, but may make performance errors. The rules of grammar—which speakers are usually unable to formulate explicitly, and may sometimes fail to follow—rule their grammatical constructions authoritatively. Even though competent speakers sometimes make performance errors, they implicitly recognize the rules of grammar as authoritative, and when their errors are pointed out they are moved to correct them. By saying that the rules of inference in a person's SBR are those that govern authoritatively, we're allowing for cases where subjects fail to follow the norms of reasoning which they implicitly recognize as the correct ones.[6] A subject who misapplies a rule of inference and arrives at a conclusion based on that reasoning does not have an internal epistemic reason to believe this conclusion.

[6]John Pollock makes a similar distinction in a slightly different context. See John Pollock, 'Epistemic Norms', in: Ernest Sosa and Jaegwon Kim, editors, *Epistemology: An Anthology*, (Malden, Mass: Blackwell, 2000), p. 218-9.

One more complication that needs to be addressed is that the rules of inference an individual follows are revisable. An individual may come to the conclusion that a certain rule of inference that authoritatively governed his thought is mistaken. For instance, an individual for whom the gambler's fallacy was an authoritative rule of probabilistic inference may come to see that he was mistaken. This individual should not have an internal epistemic reason to believe things which rely on the gambler's fallacy. Accordingly, norms of reasoning in a person's SBR should include those norms of reasoning which the subject would recognize as correct after deliberation. This clarifies what's intended by 'sound' in our formulation of internal epistemic reasons for belief. In our formula 'sound' should be understood as in accordance with the norms of reasoning that would govern the subject's thought authoritatively after due reflection.[7]

Under *epistemic norms*, I put the proper characterization of the epistemic goal for a given inquirer and the norms governing the acquisition and revision of beliefs. The epistemic goal is said to be obtaining truth and avoiding error. These two goals can pull in opposite directions.[8] Having no beliefs

[7]There is a threat of circularity here that should be acknowledged. The problem is that the revision of the rules of inference ought to be a reasoned process, but rules of inference figure in the account of reasons. One appealing way out is to claim that there is a set of norms of inference that are immune from revision which governs the revision process. As Nagel puts it: 'Not everything can be revised, because something must be used to determine whether a revision is warranted-even if the proposition at issue is a very fundamental one...No doubt, as Quine says, "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body"—but the board of directors can't be fired.' Thomas Nagel, *The Last Word*, (New York: Oxford University Press, 1997), p. 65. See also Jerrold J Katz, *Realistic Rationalism*, (Cambridge, Mass: MIT Press, 1998), p. 72-4.

[8]Keith Lehrer, *Theory of Knowledge*, 2nd edition. (Boulder, Colo: Westview Press, 2000), p. 145; Richard Foley, *Intellectual Trust in Oneself and Others*, (Cambridge: Cambridge University Press, 2001), p. 50-3.

would ensure avoiding error. However, the goal of obtaining truths would be ill-served by this policy. We need to take the risk of making errors in order to have true beliefs, and sometimes we need to knowingly accept a set of beliefs even though we know that at least one of them is false. The epistemic goal of an agent can be characterized as maximizing an epistemic value function akin to a utility function. The epistemic value function gives a certain utility to accepting a true proposition, and a negative utility to accepting a false proposition.[9] This function governs whether a subject will have a belief as to p or not-p. It does not determine whether the subject should believe p or not-p. That requires another set of norms governing the acquisition and revision of beliefs.

We can extend the value function analogy to account for these norms governing the revision of beliefs. Each inconsistency points to the existence of falsehood in our system of beliefs, and thereby to a disvalue. Each belief is an asset. The values of beliefs vary in terms of the level of confidence we have in them, and their connectedness. We want to maximize the value of our assets and minimize our costs. Faced with two inconsistent beliefs, which have equal levels of connectedness, we will discard the one that has less credibility. Faced with two inconsistent beliefs which have equal levels of credibility, we will discard the one that has less connectedness, because discarding it will require less revision in our belief system. And sometimes, we will tolerate the inconsistency, because that's the policy that serves our epistemic aims best. It's worth emphasizing that our account doesn't deny that there may

[9]The utility function analogy is due to Keith Lehrer; Lehrer, *Theory*, p. 145-6.

be beliefs that enjoy such high levels of credibility or connectedness that we consider them fixed.

What we have internal epistemic reasons to believe is not available to us right away. Reasoning, we might say, is determining what we have internal epistemic reasons to believe. We may be mistaken about what we have internal epistemic reasons to believe. I may, for instance, fail to correctly follow the norms of reasoning I take as authoritative. Furthermore, what we have internal epistemic reasons to believe may sometimes be inaccessible to us. I may be unable to find out the result of a very long multiplication if I lack the means to write down the numbers, even though I have internal epistemic reasons to believe the correct result. A subject who believes all that she has internal epistemic reasons to believe, and believes them for the right reasons, cannot change her mind about them.[10] (Needless to say, this ideal is impossible to reach for humans.) If these beliefs are mistaken she cannot come to correct them, without making, by her own lights, a mistake such as misapplying a rule of inference.

This account might seem to sever the connection between the believer and the world in an objectionable way. This is not the case. From an outsider's perspective, another individual's internal epistemic reasons for her beliefs are based on her beliefs. However, from the first person perspective, a person's

[10]How should the arrival of new information be treated? I think that we should say that what can count as new information for the subject falls under the set of beliefs a subject has epistemic reason to believe. Let's say I see a new building in Oxford. Supposing that I believe that sense perception is reliable, I have a reason to believe that there's a new building in Oxford. The belief is one in the set of beliefs I had reason to believe, though, of course, I didn't have reason to believe it before seeing the new building.

reasons for beliefs are based on facts. From the outside perspective, I can say of S, she has an internal epistemic reason to believe c, because she believes $p_1...p_n$ and accepts the norm of reasoning r, which together entail c. From the first person perspective, one reasons from the *fact* that $p_1...p_n$ to c by relying on the valid norm of reasoning r.

To put the point in another way, believing something is believing it to be true. So from the first person perspective, when one believes that p, one takes oneself to have an external epistemic reason to believe it.[11] Viewed from the first person perspective, the reason for my belief that the door to my room is open is the fact that the door to my room is open, or the facts that entail this, and not my belief that the door to my room is open.

I have refrained from formulating my discussion in terms of justified beliefs for two reasons. Firstly, justification seems to be somehow connected to truth, or truth conduciveness. However, being guided by internal epistemic reasons does not ensure arriving at the truth.[12] A cognitive agent might believe everything that she has internal epistemic reasons to believe, but be fundamentally mistaken as is the case in brain in a vat scenarios or in the evil demon scenario we began our discussion of internal epistemic reasons for belief with. Secondly, and relatedly, formulating the discussion in terms

[11] We, of course, recognize that it is possible that external epistemic reasons and internal epistemic reasons might diverge. However, we can't step out of our shoes to confirm that they don't.

[12] For instance, based on anecdotal evidence and evidence from controlled studies, Stich argues that we should reject reflective equilibrium as an account of justification, because people hold patently wrong inference rules in reflective equilibrium. See Stephen Stich, 'Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity', *Synthese*, 74 March (1988):3.

of justified beliefs would require taking a position on issues in epistemology, particularly in the internalism-externalism debate, which we don't need to be concerned with. Internal and external epistemic reasons seem to me to be coherent and informative notions. We can learn from them without entering into debates about justified belief.

I would also like to emphasize that this account is not intended to be an account of epistemic rationality. For our account of reflective equilibrium, we can remain silent about what epistemic rationality is. The relationship between internal epistemic reasons and epistemic rationality is not straightforward. I shall do no more than mention some difficulties. For instance, epistemic rationality can't be said to consist of believing everything one has internal epistemic reasons to believe, because that would be too demanding and cognitively impossible. Furthermore, epistemic rationality needs to be sensitive to one's practical interests. While a subject might have internal epistemic reasons to be believe what current astrophysics tells us about the physics of the universe, the subject could lack practical reasons to find out about them. Epistemic rationality can't also be not denying what one has epistemic reason to believe. I might have internal epistemic reason to believe that p, but that may be established through a long and complex chain of inferences. And p may seem prima facie implausible. So, this requirement would also be too demanding. We might want to say that one displays epistemic irrationality when one fails to believe p, when it should be fairly easy to establish that one has an internal epistemic reason to believe p. But this would not work, because one might mistakenly think that there are counter-

vailing reasons, and to see that those reasons are mistaken could be difficult.

2.3 From Internal Epistemic Reasons to Reflective Equilibrium

Internal epistemic reasons refer to the subjective belief and reasoning set of an individual. By filling out certain elements of the subjective belief and reasoning set of a typical moral inquirer we can arrive at a general picture of what she has internal epistemic reasons to believe. By a typical moral inquirer, I intend someone whose belief structure is widely shared and who has taken up the task of carrying out detailed reasoning about morality. I take my characterization of her to be accurate of most moral philosophers.

Our typical moral inquirer has moral convictions at different levels of generality and she has different degrees of confidence in them. Some of the moral principles she finds plausible are higher level principles like the principle of utilitarianism, or the principle 'All people are entitled to equal treatment'. Some of her principles are lower level principles such as 'Lying is wrong'. And some of her beliefs pertain to specific cases. She is confident of her moral beliefs to different degrees. Some of her beliefs are ambiguous and open to different interpretations whereas some are quite clearly spelled out. She realizes that there are contradictions between her moral beliefs. She believes that some of them may be mistaken. But she is not in the grip of sceptical doubts. She also has other beliefs in other domains such as religion, metaphysics, rationality, which can have implications for morality.

Let us approach Rawls's account of reflective equilibrium with this characterization of the typical moral inquirer and our account of internal epistemic reasons. In 'Outline of a Decision Procedure for Ethics' Rawls describes his project as discovering principles that explicate the considered judgements of competent judges. These principles would be 'such that, if any competent man were to apply them intelligently and consistently to the same cases under review, his judgements, made systematically nonintuitive by the explicit and conscious use of the principles, would be, nevertheless, identical, case by case, with the considered judgements of the group of competent judges'.[13] The process of discovering moral principles, on this account, is inductive. We begin by individual judgements. We then develop lower-level principles which account for these judgements. Finally, we seek higher-level principles which bring together our lower-level principles. This cannot be the whole picture, because, as Rawls came to see by the time of *Theory*, we have moral beliefs at different levels of generality that we invest some confidence in.

Once we acknowledge that a typical moral inquirer holds moral beliefs at different levels of generality, and adopt the above characterization of the typical moral inquirer and the account of internal epistemic reasons in the previous section, how should a typical moral inquirer reason about moral questions? Just focusing on the higher level principles and discarding lower level principles and intuitions would not make sense. She takes those beliefs to enjoy some credibility. Focusing on intuitions and ignoring the higher level principles also does not make sense, because she takes these principles

[13]John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 55.

to be likely to be true as well. So, in cases of conflict between higher-level principles, lower-level principles and intuitions, she shouldn't privilege any category.

When the principles we already hold or those we have discovered conflict with our evaluative judgements, we need to make revisions. In such cases, we may revise some of our judgements in light of these principles or we may revise our principles. At each point of conflict between our considered judgements and principles, we face a choice. We may either revise our considered judgements or the principle we have formulated, or already hold, depending on which of these we find to be more credible on due reflection. This process of revision is guided by our SBR. Through this process of revision from both ends, we approach what Rawls has called 'reflective equilibrium'.[14] Theorizing is a dynamic process where our initial beliefs are subject to revision. Accordingly, the theory we are pursuing is not one that can account for most of our considered judgements, but the one that accounts for them in reflective equilibrium. At end of theorizing we end up with a set of beliefs different from the one we began with and of which we are more confident.

It may be the case that our other philosophical beliefs and our beliefs about the natural world have a bearing on our moral beliefs. If that is the case, those beliefs should also be included in our reasoning about moral questions. In other words, we should pursue what Rawls refers to as wide reflective equilibrium. We should be careful about our reason for pursuing

[14]John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 18.

wide reflective equilibrium. Our reason isn't that the set of beliefs introduced in wide reflective equilibrium are more reliable or more objective than our moral beliefs. We aren't hoping that our moral beliefs will gain a more respectable status by rubbing shoulders with these more respectable beliefs. Rather, we are including these beliefs, because they bear on our moral beliefs. We are bringing them into the picture, because not doing this would be failing to find out things we have reason to believe.

In short, the typical moral inquirer should reason in the way Rawls's method of reflective equilibrium suggests she should. She should take her current beliefs as her starting point and make revisions as required by her SBR. She should go back and forth between her judgements at different levels of generality, and revise them according to the level of confidence she has in them to figure out what she has the strongest internal epistemic reason to believe. Rawls's methodology of reflective equilibrium is simply spelling out what a typical moral inquirer will do to figure out what she has the strongest internal epistemic reason to believe.

Theorizing, it should be noted, has an explanatory goal as well. In addition to arriving at judgements and principles we are confident of, we also want a deeper explanation of the moral beliefs we hold. As Sidgwick points out, even when we have accepted principles which formulate our beliefs, and we do not doubt their truth, we want a 'deeper explanation of *why* it is so'.[15] Different moral theories such as contractualism and utilitarianism provide

[15]Henry Sidgwick, *The Methods of Ethics*, 2nd edition. (London: Macmillan, 1877), p. 91, emphasis in the original.

us both with principles that they claim match our judgements and an explanation of these judgements.[16] Sidgwick, for instance, puts forward the explanatory hypothesis that common sense morality is a 'machinery of rules, habits, and sentiments, roughly and generally but not precisely or completely adapted to the production of the greatest possible happiness for sentient beings generally'.[17] Rawls's theory can (roughly) be interpreted as saying that our intuitions about and lower-level principles of distributive justice are accounted for by the principles which are justifiable to each individual. The explanation we seek is one which also has justificatory force. The principle or set of principles that account for our beliefs should have independent appeal. Once we are aware of these principles we should be able to endorse them and our moral beliefs that follow from them. By a moral theory, then, I understand a single principle or a set of higher-level principles which account for other lower-level principles and intuitive judgements, and have explanatory and justificatory power.

At the end of this process, the typical inquirer has two reasons to be more confident of the moral beliefs she has arrived at. Firstly, she has uncovered and removed existing inconsistencies. So there is no obvious sign that some of her beliefs are false. Secondly, principles act as 'transmission devices for *probability* or *support*' like lawlike statements in the natural sciences.[18] When one is able to subsume several intuitive judgements under a principle that

[16]T. M. Scanlon, 'Contractualism and Utilitarianism', in: Amartya Sen and Bernard Williams, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982), p. 108.

[17]Sidgwick, *The Methods of Ethics*, p. 435

[18]Robert Nozick, *The Nature of Rationality*, (Princeton: Princeton University Press, 1993), p. 5, emphasis in the original.

one finds plausible, one's confidence in both the principles and the intuitive judgements increases. The following example from Chisholm nicely illustrates Nozick's point.[19] Suppose you see a cat on the roof on Monday. Then on the following three days you see the same cat. This gives you a good reason to believe that there is a cat on the roof almost everyday. Now this further claim gives you reason to be more confident of your belief that you saw a cat on the roof on Monday. Your initial reason for thinking that there was a cat on the roof on Monday was your observation of a cat. Now that evidence is coupled with the general observation that there is a cat on the roof almost everyday. You now have two reasons to believe there was a cat on the roof on Monday. We should note that in the case of reasoning about morality the relationship of mutual support will tend to be stronger, because the higher level principle, which is analogous to the belief that there seems to be a cat on the roof almost everyday, will enjoy some plausibility independent of the intuitions it accounts for.[20]

2.4 Implications of This Account of Reflective Equilibrium

On my account, Rawls's account of reflective equilibrium is an account of the reasoning process of typical moral inquirers. It is not aimed to justify the reliance on moral intuitions. All of the moral principles and intuitions

[19]Roderick M. Chisholm, *Theory of Knowledge*, 2nd edition. (Englewood Cliffs: Prentice-Hall, 1977), p. 83-4.

[20]The analogy between moral reasoning and Chisholm's example would be more exact if in addition to our observations, we were told by a person we believe to be reliable that there was a cat on the roof almost everyday.

which play a part in the pursuit of reflective equilibrium are ones that enjoy prima facie credibility for the typical moral inquirer. It is also not in competition with theories of justification. It doesn't assert that coherence amongst one's beliefs is *sufficient* for epistemic justification. It should therefore be acceptable to proponents of different views on justification. Rawls's account of reflective equilibrium is just based on the accurate observation that we have moral beliefs at different levels of generality and there is no reason to privilege one level at expense of others, and that our SBR is all that we have to go on.

Even though reflective equilibrium is not an account of epistemic justification, it is enlightening. Consider how it can help us view sceptical challenges. Suppose we have a deductively valid philosophical argument for the conclusion that 'All your moral beliefs are mistaken'. The argument has premises, $p_1...p_n$, which are intuitively plausible. When we are presented with this argument, our mind will naturally be drawn to $p_1...p_n$ and their plausibility. However, the account of reflective equilibrium will remind us that we should take a wider view of the argument. We should ask how it relates to our other beliefs. Once we cast our net wider, we will see how despite their intuitive plausibility, there must be something wrong with the premises. For the argument implies that your belief that 'Torturing babies for fun is wrong' is also mistaken. Reflective equilibrium reminds us that the premises of the argument and the other beliefs it contradicts with come from the same stock. Note that reflective equilibrium does not help us refute the sceptical argument. It just makes it easier for us to see why the sceptic can't win us

over.

The same lesson applies to metaethical theories. From our perspective metaethical theories and ethical theories are in the same boat. Therefore, a factor in our evaluation of metaethical theories will be how they relate to our first-order moral theories. Metaethics seeks to provide an account of what we do when we engage in moral argument. It deals with questions of meaning, the relationship between moral beliefs and motivation, metaphysical issues such as whether moral facts exist, and if they do what sorts of facts they are, and how we can come to know them. Some metaethical theories leave our first-order moral discourse as it is. Our acceptance of these metaethical theories does not require revision of our moral beliefs. Other metaethical accounts, however, make claims on first-order moral discourse and favour a specific first-order theory.[21] Once we accept these metaethical theories, we need to revise our moral beliefs or discard them altogether. In such cases, we are faced with the choice of revising our metaethical or first-order beliefs.

Suppose on one metaethical theory, *M*, moral facts are natural facts, and this metaethical theory rules out all moral theories except utilitarianism. Then, whether a person accepts *M* or not will depend not only on how *M* accounts for her convictions at the metaethical level, but also on how it relates to her first-order convictions. If the acceptance of *M* will, on the whole, be a loss to her in terms outlined on page 23, then she has a stronger reason to reject *M*. In this case, she would either have to revise some of the

[21] An example of such a metaethical theory is Peter Railton's realism. Peter Railton, 'Moral Realism', *The Philosophical Review*, 95 (1986):2.

beliefs she had which lead her to accept M, or she may seek an alternative to M. This does not mean that our ethical commitments should govern our choice of metaethical theories.[22] When considering competing metaethical theories that leave moral discourse untouched, the choice of theory is in no way influenced by our ethical views. However, if a metaethical theory makes claims which conflict with a person's first-order moral claims, then it would need to have the necessary weight to displace that person's moral convictions. And this will be unlikely. Metaethical theorizing is driven by beliefs about knowledge, particularly about how we acquire knowledge, the nature of facts and truth, and motivation. A typical moral inquirer is much less confident in these beliefs than first-order moral claims. For instance, some metaethical theories rely on theories of knowledge to argue for moral nihilism. But surely, we are more confident of our first order moral beliefs than some abstract theory of knowledge. Reflective equilibrium reminds us of this fact.

2.5 Conclusion

By framing reflective equilibrium as an account of the reasoning process of typical moral inquirers, we are able to avoid unnecessary ambitions that are sometimes attributed to reflective equilibrium. Firstly, on this account, reflective equilibrium is not an account of epistemic justification. It is perfectly possible that a person who has reached reflective equilibrium has un-

[22]Dworkin makes a similar argument to the one offered here, so I should emphasize one major difference. Dworkin argues that *all* metaethical theories make first-order claims. My argument here neither affirms nor denies this claim. See Ronald Dworkin, 'Objectivity and Truth: You'd Better Believe it', *Philosophy and Public Affairs*, 25 (1996):2.

justified beliefs. Secondly, this account of reflective equilibrium takes the typical moral inquirer and her beliefs as given. It doesn't set itself the task of demonstrating that one's moral intuitions are reliable. Its only claim is that granted that our beliefs and initial confidence levels were accurate, at the end of theorizing we will be in a better epistemic position. I argued that someone who follows the method of reflective equilibrium will be able to resist sceptical challenges, but this does not amount to refuting them. The modesty of reflective equilibrium's claims helps explain why it is so popular.

We might ask whether the account presented here is descriptive or normative. The account of the typical moral inquirer is descriptive. However, I'm not sure about how the rest of the account should be characterized. Is it an account of how people reason, or how people *ought* to reason? The account seems descriptively accurate to me. It also seems to embody some normative commitments. For instance, someone who uncovers two inconsistent beliefs with equal levels of connectedness, but different levels of credibility, and discards the belief that she finds more credible would be doing something wrong. But it is difficult to imagine someone doing this. If we came across this person, we would interpret her as judging the discarded belief as less credible. The deeper problem, here, is the following. An account of how people ought to reason would itself be an instance of reasoning. If it says that people fail to reason as they ought to, then it would undermine itself.[23] One would be saying that by reasoning in a way I ought not to have, I have arrived at an account of how I ought to reason. In short, the normative and the descriptive

[23]Note that I'm talking about a blanket statement about reasoning rather than a statement about some rules of inference etc.

aspects of an account of reasoning are inextricably intertwined. Nevertheless, we can say this much. Since this model involves some level of idealization, it has normative elements. It is possible that we fail to follow it on some occasions due to our cognitive shortcomings, but the ideal is one we closely approximate.

In this chapter I've set out the framework that informs this thesis. The emphasis on the perspective of the individual inquirer informs the discussion of the epistemic significance of moral disagreement we shall look at in the next chapter. The conclusion of section 2.4 on the status of metaethical theories in relation to first-order theories informs the discussion of constructivism in chapter 4. The framework of moral theorizing forms the background of the discussion of Cohen's critique in chapter 5.

Chapter 3

The Epistemic Significance of Moral Disagreement

3.1 Introduction

The account of reflective equilibrium presented in the previous chapter focused entirely on the perspective of the individual inquirer. It adopted the viewpoint of an individual inquirer, and set out an account of her reasoning. In this chapter, we bring into the picture other inquirers, and ask what is the proper role of their beliefs in our reasoning and how interpersonal justification works. In particular, I want to examine whether there's a close connection between what we are rationally entitled to believe, and what we can justify to others. The significance of this question for our project arises in the following way. In the next chapter, I will argue that we should attribute to Rawls two distinct constructivist projects. The first project adopts the first-person perspective and aims to arrive at an account of justice that is in

reflective equilibrium for that person. The second project adopts an interpersonal perspective, and aims to arrive at principles that can be legitimately imposed on adherents of different comprehensive doctrines. My argument in the next chapter will be that these two projects operate with different constraints and different resources. In order to understand the constraints on the second project, and the resources at its disposal, we need an account of interpersonal justification. In order to maintain that the two projects are distinct, we need to establish that there isn't a close link between what we can justify to others and what we are rationally entitled to believe. In other words, we need to show that the following requirement is false, or at least not applicable in most cases.

> *The congruence requirement* One should believe only those moral principles that one can justify to others.[1]

The most plausible and natural way to argue for the congruence requirement is by arguing that if other rational people disagree with us on a philosophical question, this is epistemically significant. The main idea behind this is succinctly expressed by Sidgwick:

> 'I suppose that the conflict in most cases [of philosophical disagreement] concerns intuitions–what is self-evident to one mind is not so to another. It is obvious that in any such conflict there must be error on one side or the other, or on both. The natural man will often decide unhesitatingly that the error is on the other side. But it is manifest that a philosophic mind cannot do this,

[1] 'Justify' is ambiguous. Here I am using it in an epistemic and not moral sense.

> unless it can prove independently that the conflicting intuitor has an inferior faculty of envisaging truth in general or this kind of truth; one who cannot do this must reasonably submit to a loss of confidence in any intuition of his own that thus is found to conflict with another's.'[2]

If Sidgwick is correct, then we have good reason to honour the congruence requirement. If we interpret 'loss of confidence' to mean withholding judgement on intuitions and principles which are matters of contention, we can only reason on the basis of principles and intuitive judgements which we share with other people.[3] As a result the moral principles we are entitled to believe will be ones which we can justify to others.

Here is how I shall proceed. I begin by presenting Rawls's account of interpersonal justification, which I take to be an accurate account of how interpersonal justification works (3.2). With that account in place, I argue that honouring the congruence requirement would require us to (a) hold weaker views than we would in the absence of the congruence requirement, and (b) to hold views that would seem incorrect to us in the absence of the congruence requirement (3.3). Showing that the congruence requirement has these consequences is not enough to refute it. For if Sidgwick is correct, it is the views which we can justify to others that we should hold.

[2]Henry Sidgwick, *Essays on Ethics and Method*, (Oxford: Clarendon Press, 2000), p. 168. I came across this quote in Ralph Wedgwood's discussion of similar questions in Ralph Wedgwood, *The Nature of Normativity*, (Oxford: Clarendon Press, 2007), p. 262.

[3]I adopt this strong reading of Sidgwick for two reasons. Firstly, a weaker reading of 'less confidence' does not support the congruence requirement. Secondly, it is the strong reading that has found support in the literature.

The epistemic significance of disagreement, particularly amongst epistemic peers, has been a lively topic in recent years, and positions similar to Sidgwick's have found support. I turn to this debate to examine whether conciliatory responses to disagreement like Sidgwick's, which would support the congruence requirement, are always appropriate (3.4). I begin by examining some ways to respond to peer disagreement, which have been accommodated by conciliatory views (3.5). I then move on to the Independence principle, which is central to conciliatory views. To evaluate this principle, I survey some of the ways in which we determine others to be reliable (3.6). I argue that some ways of judging another person as an epistemic peer justify violations of Independence (3.7). I then examine whether the number of people we disagree with makes a difference (3.8). Finally, I argue that in a wide range of instances of philosophical disagreements, including those about morality and politics, the conciliatory response isn't required (3.9).

3.2 Rawls on Interpersonal Justification

Rawls offers an account of how interpersonal justification works:

> '... [J]ustification is argument addressed to those who disagree with us, or to ourselves when we are of two minds... Being designed to reconcile by reason, justification proceeds from what all parties to the discussion hold in common. Ideally, to justify a conception of justice to someone is to give him proof of its principles from premises that we both accept, these principles having in turn consequences that match our considered judgements. Thus

> mere proof is not justification. A proof simply displays logical relations between propositions. But proofs become justification once the starting points are mutually recognized, or the conclusions so comprehensive and compelling as to persuade us of the soundness of the conception expressed by their premises.'[4]

There is an ambiguity in Rawls's presentation, which we need to address right away. The initial condition Rawls introduces for justification is that we produce an argument for it from premises held in common to conclusions which are held in reflective equilibrium. This is the condition expressed in the sentence beginning with 'Ideally'. However, in the last sentence which I've quoted, Rawls states a more relaxed condition. We are to provide *either* a deductively valid argument from premises held in common, or the argument we provide will generate a conclusion 'so comprehensive and compelling' that we'll conclude that the premises have to be true.

Rawls's weaker requirement can easily fall short of creating agreement. Consider Rawls's first suggestion of arguing from shared premises to a conclusion. As Lycan aptly puts it 'a deductive "proof" can be no more than an invitation to compare plausibility'.[5] When faced with a deductively valid argument that relies on premises we accept, our acceptance of the conclusion of this argument is not inevitable. The acceptance of the premises does not entail our being certain of them. If the conclusion strikes us as implausi-

[4]John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 508.

[5]William G. Lycan, 'Moore Against The New Skeptics', *Philosophical Studies*, 103 (2001):1, p. 39.

ble, we have the choice of either embracing the implausible conclusion, or dropping one of the premises that we initially accepted, which entailed the conclusion we find implausible.

Now, let us consider Rawls's second suggestion: justifying the premises by the comprehensive and compelling nature of the conclusion. The difficulty here is that there will be several arguments that can support the conclusion in question. Singling out the premises offered will be a difficult task. Firstly, the premises offered by the other person may seem implausible or even wrong to us. And, given the presence of other premises which support the same conclusion, we will be drawn to the other premises. Secondly, showing a set of premises to be the ones to be picked is a complex task. The conclusion may seem 'comprehensive and compelling', but there will be several other candidates that can claim the same virtues. These alternatives may seem better at conserving the other person's beliefs, or more easily derivable from their beliefs.

Given the shortcomings of Rawls's weaker requirement, we need to turn to his stronger requirement. That requirement is that the premises of the argument offered are held in common, and the consequences it generates 'match our considered judgements'. There is a question here about the extent of the match with our considered judgements. Does Rawls require that the principles match all of the considered judgements of both individuals or only those considered judgements that they hold in common? The rationale for excluding premises that are not shared would apply to the considered judgements

that are not shared. Accordingly, I interpret Rawls as requiring that only the considered judgements which are held in common can be relied on in interpersonal justification.[6] So, I can offer you an interpersonal justification of a principle P by showing that (a) P follows from premises that we both hold; and (b) P matches considered judgements that we share.

This account of interpersonal justification coupled with Sidgwick's advice-on the strong reading which I've adopted here-would give us the congruence requirement. By following Sidgwick's requirement, we exclude principles and intuitions we disagree on from our reasoning. For the fact that others deny them is a reason to discount these intuitions and principles. And Rawls's account shows how relying only on those principles and intuitions we share produces interpersonal agreement.

3.3 The Implications of the Congruence Requirement

The problem with the congruence requirement given this picture of interpersonal justification is that arguments which rely only on the shared premises and shared considered judgements will fall short of what individuals are entitled to believe. The congruence requirement narrows down the basis our moral theory can build on. A theory that covers most of our considered con-

[6]Note that Rawls is not suggesting that any justification will do. Justification ought to be based on what we both believe to be the case. For instance, if I'm an unbeliever, and you're a devout Christian, and I convince you of my views on the basis of the Bible, this would not count as justification in the relevant sense.

victions is a better theory from our perspective than one which builds upon such a narrow base. There are two distinct problems. The theory that builds only on shared beliefs, will sometimes be weaker than the theory that takes in all of one's beliefs. I shall call this *the weak view problem*. Sometimes, the theory which is built only on shared beliefs, will be false according to the the theory that takes in all of one's beliefs. I shall call this the *false views problem*.

Let us first look at the disagreement between an egalitarian and a prioritarian for an illustration of the weak views problem. According to *telic egalitarians*, 'It is in itself bad if some people are worse off than others'.[7] According to *prioritarians*, 'Benefiting people matters more the worse off these people are'.[8] As Parfit points out, these two views often coincide on their judgements of cases. So, they can be offered as possible explanations of similar intuitions. Nevertheless, they are distinct views, and they offer different verdicts on the following case:

> 'Suppose that, in some natural disaster, those who are better off lose all their extra resources, and become as badly off as everyone else.'[9]

According to telic egalitarians, since inequality is removed, this change is in one respect a change for the better. According to prioritarians this is in no way a change for the better since no one's condition has improved.

[7]Derek Parfit, 'Equality and Priority', *Ratio*, 10 (1997):3, p. 204.
[8]Parfit, 'Equality and Priority', p. 213.
[9]Parfit, 'Equality and Priority', p. 210-1.

Suppose we have a telic egalitarian and a prioritarian. They share several considered judgements. Both, for instance, think that of the two outcomes

1. Half at 100 Half at 200
2. Everyone at 145

the second outcome is better. When they exclude the considered judgements they disagree on, in this case the judgement about levelling down, they can only justify a view which has narrower scope than telic egalitarianism and prioritarianism. But from the perspective of someone who has the considered judgement that in Parfit's example there is, in one respect, a change for the better, egalitarianism makes more sense. If we require him to believe only what he can justify to the prioritarian, he'll be required to hold a view which has less scope than the one it is rational from his perspective to hold.

As a second example, let us consider Sen's intersection approach to measuring inequality. There are several different proposals for the correct measurement of inequality. Sen observes that even though these measures differ in certain instances, there are also cases where they converge. Sen's proposal is to rely on the judgements these different measures agree on.[10] Sen offers the following example. Let's say we want to compare the UK, the USA, Ceylon, and Mexico in terms of inequality. Three measures of inequality –the Gini coefficient, the coefficient of variation, and the standard deviation of algorithms– produce the same rankings in this case. So we can agree that, for instance, Mexico is more unequal than the UK.

[10] Amartya Sen and James E. Foster, *On Economic Inequality*, Enl. edition. (Oxford: Clarendon Press, 1997), p. 72.

Sen's intersection approach models Rawls's picture of interpersonal justification.[11] We leave out the cases of disagreement, and build on what's held in common. Let's say the Gini coefficient is able to capture my intuitions about the extent of inequality and the coefficient of variation captures your intuitions about inequality. So, based on what we hold in common, it will produce an answer which we agree with. However, their rankings diverge in the case of India and Ceylon.[12] According to the Gini coefficient, India has less inequality than Ceylon. According to the coefficient of variation measure, Ceylon has less inequality than India. So, the intersection approach will be silent in the case of Ceylon and India. However, from our perspectives there is a definite answer on the question of whether Ceylon or India is more unequal. The rankings we can justify to each other will be incomplete. If we believe only what we can justify to another person, we will be throwing out information which from our perspective is relevant.

In both of the previous examples, it was shown that relying only on premises or considered judgements shared by all the parties to a disagreement would result in a view that had a narrower scope and left out what we thought was true. There can also be cases where relying on only the premises both parties hold in common produces results that are false by the lights of one of the parties. This is what I called the false views problem.

[11] Sen makes modest claims for this approach. Furthermore, Sen has a different motivation for putting forward the intersection approach. So, what I say here is not intended as a criticism of Sen.

[12] Sen and Foster, *Economic Inequality*, p. 73.

Here is one example which shows how this can happen when reasoning about empirical questions. Suppose you and I are wondering whether Elif eats pork. We both believe that she is from the U.K. In addition to this, I believe that she is a Muslim whereas you believe that she is not a Muslim. I have good, but defeasible, reason to believe that she will not eat pork, because most Muslims do not eat pork, and the country they are from is unlikely to change this. However, if we reason only on the basis of the premises we hold in common, then we should expect that Elif will eat pork, because most people from the U.K. do eat pork.[13]

Here is another example of the false views problem, this time involving an example of moral reasoning. Suppose you and I are considering whether to obey Creon's edict that Polyneices's body should not be buried. We both believe that we have a pro tanto reason to obey the law. In addition to this, I believe that we also have a pro tanto reason stemming from respect for the dead to bury Polyneices's body. Furthermore, I believe this over-rides the first reason. You, however, believe that we don't have such a reason, because Polyneices's conduct in the war disqualifies him from considerations of respect for the dead. If we reason only on the basis of the reasons we hold in common, then we should hold that we have a pro tanto reason of political obligation to not bury Polyneices. Since there are no countervailing reasons

[13]We might worry that in this example, we've ended up taking the rejection of 'Elif is a Muslim' to be case. That this worry is misplaced can be seen by distinguishing between the following three conditional probabilities: The probability that Elif eats pork given that she's from the U.K., the probability that Elif eats pork given that she's from the U.K. and she is a Muslim, the probability that Elif eats pork given that she's from the U.K. and she is not a Muslim. The negation of 'Elif is a Muslim' would give us the last conditional probability which is different from and higher than the first conditional probability.

requiring that we bury Polyneices that we hold in common, we should also conclude that we have conclusive reason to obey Creon's edict. The answer reached when we relied only on the beliefs we shared and the answer I would reach on my own produce different recommendations.

We can perhaps live with the requirement of believing only what we can justify to others when the only outcome of this policy is ending up with views which are narrower in scope than the ones we would have if we did not follow this policy. However, ending up with beliefs which are false by our own lights is troublesome. It should be clear from these examples that the views we end up with when we honour the congruence requirement and the views we would end up with if we don't honour it differ significantly.

Here's one way of resisting my argument. The reasoning in the Elif example is not deductively valid. It is possible for the premises to be true, but the conclusion to be false. Elif may be a Muslim from the U.K. who eats pork. In order to avoid the implications of the Elif example, we could restrict the congruence requirement so that it applies only to deductive reasoning. This restriction, however, would make the congruence requirement almost inapplicable to moral reasoning. Even though the reasoning in the Elif example is not deductively valid, it is rationally compelling, and it is a kind of reasoning that is essential to reasoning about morality.

3.4 The Peer Disagreement Debate

I have shown that accepting the congruence requirement has serious costs. This establishes a good case against the congruence requirement. However, if Sidgwick is correct we have reason to discard the beliefs others disagree with. It would be the narrower views, or the views which would have seemed wrong to us if beliefs others dispute were admissible, that we ought to hold. We should, therefore, turn to an examination of Sidgwick's principle.

The proper epistemic response to disagreement with others has been the topic of recent discussions. [14] The central question in the debate has been the following: when I believe that p and I find out that someone I consider an epistemic peer believes that not-p, or when I have high credence in p and an epistemic peer has low credence in p, what is the proper response to this fact?[15]

In cases of peer disagreement, *the conciliatory response*, the contemporary representative of Sidgwick's view, 'mandate[s] extensive revision in our opin-

[14]David Christensen, 'Epistemology of Disagreement: The Good News', *Philosophical Review*, 116 (2007):2; David Christensen, *Disagreement, Question-Begging and Epistemic Self-Criticism*, Unpublished manuscript ⟨URL: http://www.brown.edu/Departments/Philosophy/onlinepapers/christensen/Conciliationism.pdf⟩; Adam Elga, 'Reflection and Disagreement', *Nous*, 41 (2007):3; Adam Elga, *How to Disagree about How to Disagree*, Unpublished manuscript ⟨URL: http://philsci-archive.pitt.edu/archive/00003702/⟩; Thomas Kelly, 'The Epistemic Significance of Disagreement', in: John Hawthorne and Tamar Gendler Szabo, editors, *Oxford Studies in Epistemology*, Volume 1, (Oxford: Clarendon Press, 2005); Thomas Kelly, *Peer Disagreement and Higher Order Evidence*, Unpublished manuscript ⟨URL: http://www.princeton.edu/~tkelly/papers/KellyPeerDis-1.doc⟩; and the collection of articles R. Feldman and T. Warfield, editors, *Disagreement*, (Oxford: Oxford University Press, Forthcoming).

[15]Note that our question is narrower than the question of what would be rational to believe about p all things considered. This however does not mean that the answers to these two questions are wholly independent. Christensen, 'Question-Begging', p. 6.

ions'.[16] According to *the equal weight view*, which is the most commonly defended example of the conciliatory response,

> 'In cases of peer disagreement, one should give equal weight to the opinion of a peer and to one's own opinion.'[17]

According to conciliatory views, once I find out that an epistemic peer has low credence in p my credence in p should be lowered. Or, if we adopt an all-or-nothing model of belief, when I find out that an epistemic peer believes that not-p and I believe that p, I should suspend judgement on p. There are cases where the conciliatory response is intuitively compelling. Here is one such case.

> *The Horse Race* 'You and I, two equally attentive and well-sighted individuals, stand side-by-side at the finish line of a horse race. The race is extremely close. At time t0, just as the first horses cross the finish line, it looks to me as though Horse A has won the race in virtue of finishing slightly ahead of Horse B; on the other hand, it looks to you as though Horse B has won in virtue of finishing slightly ahead of Horse A. At time 1, an instant later, we discover that we disagree about which horse has won the race.'[18]

In this case, it is clear that my credence in Horse A winning should be significantly lowered. Here is another case, which supports the conciliatory response.

[16] Christensen, 'Question-Begging', p. 1. The term conciliatory is due to Elga. See Elga, 'How to Disagree', p. 1

[17] Kelly, 'Peer Disagreement and Higher Order Evidence', p. 2. I should mention that Kelly is a critic of the equal weight view.

[18] Kelly, 'Peer Disagreement and Higher Order Evidence', p. 3.

> *The Murder Investigation* There has been a murder, and Jones is the prime suspect. There are two police investigators independently working on the case. They both have the same evidence. They are equally intelligent and thorough in their analysis of the evidence. You have no reason to suspect that either of them is biased. You learn that one of them thinks Jones committed the murder and the other thinks he didn't.

What should you think? You should give equal credence to the hypotheses that Jones committed the murder and that Jones did not commit the murder, because by your lights the two police officers are equally likely to be correct. Suppose that you are, in fact, one of the police investigators. Shouldn't you reason in the same way?

The conciliatory response is further strengthened by drawing analogies with the following kind of cases.

> *The Thermometers* 'You and I are each attempting to determine the current temperature by consulting our own personal thermometers. In the past, the two thermometers have been equally reliable. At time t0, I consult my thermometer, find that it reads '68 degrees', and so immediately take up the corresponding belief. Meanwhile, you consult your thermometer, find that it reads '72 degrees', and so immediately take up that belief. At time t1, you and I compare notes and discover that our thermometers have disagreed.'[19]

[19]Kelly, 'Peer Disagreement and Higher Order Evidence', p. 4

The Horse Race and the Thermometers seem analogous. Just as it would be wrong to take the fact that one of the thermometers is *mine* as a reason to favour its reading, it would be wrong to stick to my belief in the Horse Race.

The conciliatory response might seem commonsensical. Yet, it would have radical implications in other areas of philosophy. Think of the big questions in philosophy, for instance, the question of God's existence. Suppose you think that God does not exist. Nevertheless, a lot of thoughtful and intelligent people think that God exists. The conciliatory view would require you to give up atheism and move closer to agnosticism.

Despite the intuitive appeal of the conciliatory view, sometimes when we find out that people we would consider to be epistemic peers don't share our responses this doesn't shake our confidence. We are inclined not to lose our confidence in our belief, but demote others from the status of epistemic peer because of the disagreement between us. We are inclined to take the belief they have expressed to reveal a fact not about the world but about themselves. In the next three sections, I shall argue that this response is the correct response to peer disagreement in some cases and the conciliatory response is the correct response in others depending on our grounds for having judged the other person to be an epistemic peer.

3.5 Preliminary responses to Peer Disagreement and the Independence Principle

I shall define a person's *epistemic peer* as someone who is as reliable as that person in a certain domain.[20] By this I mean that epistemic peers are equally likely to arrive at true beliefs in a given domain. More formally, Person A and Person B are epistemic peers on the question whether p or not-p, if and only if, when p is the case, the probability of A judging that p is equal to the probability of B judging that p, and when not-p is the case, the probability of A judging not-p is equal to the probability of B judging not-p.

The question I'm interested in is what the proper response to disagreement with someone I *consider* to be an epistemic peer is. This way of formulating the question requires us to acknowledge several possible sources of error on my part, which need not concern us if we take an external point of view, and ask what to do when we disagree with someone who *is* as reliable as we are. When I judge that p and someone I consider an epistemic peer judges not-p, there are three beliefs which are candidates for revision: my belief that p (or my credence n in p), my belief that the person I'm disagreeing with is an epistemic peer, and my belief that we are in fact disagreeing. When we take up an external view point, the only candidate for revision will be my belief that p.

[20]There are several different definitions of epistemic peer in the literature. See for instance Christensen, 'Epistemology of Disagreement', p. 188 Elga, 'Reflection and Disagreement', p. 499 Kelly, 'Epistemic Significance', p. 174-5. In 3.6, I shall say more to defend my definition of epistemic peer.

Let us begin with the last candidate for revision. When I say that p I know that I'm expressing what I take to be the truth. However, I don't know with the same certainty that you are also expressing what you take to be the truth. There is an important asymmetry between us regarding our knowledge about your intentions.[21] When you say that not-p, you may be lying, or joking. So, my judgement that we are in fact disagreeing may be in error. In certain cases, believing that we are not actually disagreeing may be more sensible than believing that either my judgement that p or my judgement that you are an epistemic peer is in error.

My belief that the person with whom I'm disagreeing is an epistemic peer can be revised in two ways. Firstly, I might decide that even though you are an epistemic peer, on this occasion you are not. There may be an immediately available explanation of why, on this specific instance, you are more likely to be mistaken than I am. I might, for instance, think that on this occasion your cognitive faculties are not functioning properly, because you're inebriated.

Here we should note another asymmetry between our respective positions. I am in a position to monitor the functioning of my cognitive faculties more closely than I am in a position to monitor the functioning of your cognitive faculties. As Lackey points out:

> 'I may, for instance, know about myself that I am not currently suffering from depression, or not experiencing side effects from prescribed medication, or not exhausted, whereas I may not know

[21] Christen also makes this point. See Christensen, 'Question-Begging', p. 14.

that all of this is true of you.'[22]

It is as if I can examine my thermometer and rule out the possibility of its being broken in an obvious way, but can't rule out that possibility with regard to your thermometer.[23]

Secondly, I might decide that you were not after all an epistemic peer, because you think not-p when p is the case. Those who are in favour of a conciliatory response to peer disagreement have been willing to accept the previous proposals but draw the line here. They hold that the previous responses are in agreement with the spirit of their view. What is crucial for the conciliatory response is that our response to peer disagreement respect what Christensen calls Independence:

> '*Independence*: In evaluating the epistemic credentials of another's expressed belief about P, in order to determine how (or whether) to modify my own belief about P, I should do so in a way that doesn't rely on the reasoning behind my initial belief about P.'[24]

This principle does a good job of capturing what would be odd in deciding that your thermometer was unreliable, in the Thermometer Case, by looking

[22] Jennifer Lackey, *A Justificationist View of Disagreement's Epistemic Significance*, Unpublished manuscript ⟨URL: http://faculty.wcas.northwestern.edu/~jal788/documents/JustificationistDisagreement-SE.pdf⟩, p. 16.

[23] The asymmetry between our respective positions may also give me reason to demote *myself* from the status of epistemic peerhood. For instance, on a given question I may be prone to self-deception. Or I may know that I'm tired, and so I'm more likely to be mistaken. I'm grateful to David Miller for reminding me of this point.

[24] Christensen, 'Question-Begging', p. 2. It's worth pointing out that Sidgwick affirms a similar principle in the passage I've quoted on page 39.

at my thermometer. It seems that it would be question begging on my part to reason that your thermometer was unreliable because it disagreed with mine.

I shall argue that Independence is a principle we should often, but not always, respect. More precisely, in cases of peer disagreement, we should not respect Independence when (a) our judgement that another is an epistemic peer can only be established by a route which does not respect Independence, or (b) when we put less trust in the reasons we relied on in judging the other to be an epistemic peer than the reasons we relied on in judging that p–the statement our epistemic peer disagrees with. In order to illustrate this I need to first examine some of the ways we can determine that someone is reliable.

3.6 Judgements of Reliability

How do the beliefs of others enter into my deliberations about what to *believe*?[25] Others' beliefs enter into our deliberations in the way that other evidence does: as indicators of what is the case.[26] You call me from Istanbul. You tell me that the weather there is cold. I come to believe that the weather in Istanbul is cold. I've set up a webcam that shows a thermometer in Istanbul. I look at the webcam. It reads 4 ℃. I come to believe that the

[25]I emphasize 'believe', because others' beliefs may also enter into my practical deliberations. For instance, in order to guess what you'll do I will rely on what I take you to believe. I won't be considering these kinds of deliberations. Here, I shall also not address the role of others' beliefs in guessing what other beliefs they have or will have, or how they will act.

[26]The same point is made in Kelly, 'Peer Disagreement and Higher Order Evidence', p. 21-2.

weather in Istanbul is cold. Your belief functions in the same way as reading the thermometer. I take both your belief and the thermometer as indicators of the temperature in Istanbul. Insofar as I consider you a reliable indicator of the temperature, I will give weight to your beliefs about the temperature.

How do I decide whether you're reliable or not? In many cases we assume that people are reliable indicators of various states of the world. Stopping to determine the reliability of others in every instance would have a paralyzing effect on our deliberations.[27] In many domains, we treat others as reliable unless proven unreliable. We can call this *presumed reliability*. When I ask you about the weather in Istanbul, I will, naturally, assume that (a) you will answer me truthfully; and (b) your answer is likely to be correct. Under normal circumstances, I wouldn't try to determine whether you are reliable. Presumed reliability is a very rough judgement. We presume that people are reliable, not that they are equally reliable. For that reason it can't be used to establish that two people are epistemic peers.

Not all judgements of reliability are presumed. There's also, what I will call *inferred reliability* where we don't presume that a person is reliable but determine that she is so. There are two main ways of determining someone's reliability. Firstly, I can judge that they are reliable by examining their track record. Secondly, background theories I am committed to may suggest that they are reliable.

[27]For a discussion of our reliance of others as sources of information and what I here call presumed reliability see C. A. J Coady, *Testimony: A Philosophical Study*, (Oxford: Clarendon Press, 1992).

Suppose I wonder whether you're reliable at doing sums in your head. I give you a set of numbers. You add them up in your head. I do the sums on paper, and check my results twice. We repeat this procedure several times. Insofar as you come up with the results I come up with, I'll conclude that you are reliable. In this case, I have what I take to be a very reliable way of checking the results you've produced. Note that in judging whether you're reliable at doing mathematical operations in your head, I am not assuming that I am reliable at doing mathematical operations in my head. I'm assuming that the method of doing the sums on paper and checking the results is a reliable method. I could examine whether I'm reliable by using the same method and come to decide that we are equally reliable.

We can also make judgements about the reliability of individuals who are in a position to be more reliable than us. For instance, I can judge how reliable a meteorologist is despite my ignorance about meteorology. I can simply check a given meteorologist's track record to determine whether she is reliable. Even though today I lack information to evaluate a meteorologist's prediction about tomorrow's weather, tomorrow I'll be in a position to evaluate her prediction.

In both of the previous examples, in judging someone else's reliability we took up an epistemically more secure position than them. Doing sums on paper and checking it twice is a more reliable way of doing sums than doing them in one's head. Finding out about the weather at t_1 by checking the weather at t_1 is more reliable than predicting at t_0 what the weather at t_1

will be. I shall refer to these kinds of judgements as *track record judgements from a more secure epistemic position.*

Sometimes we don't have the opportunity to judge someone's reliability by taking up a more secure epistemic position than them. In those cases we have to decide whether they are reliable by comparing what they take to be the case with what we take to be the case without taking up a more secure epistemic position. I shall call these judgements *track record judgements from the ground level.*

Here's an example of a track record judgement from the ground level. Suppose I'm an art historian. I hear in the news that another art historian claims that a drawing attributed to Michaelangelo is a forgery. Another art historian denies this. I can't study the drawing myself. However, as luck would have it, there are several drawings attributed to Michaelangelo where I am, and both art historians have studied them. There's a record of which ones they thought were genuine and which ones they thought were forgeries. In addition to this, I know that none of these studies involved any technological instruments. They were just done by close observation. Lacking instruments that would enable me to make track record judgments from a more secure epistemic position, I can look at the drawings and compare judgements they have made with mine to determine which art historian to trust. If one of them seems consistently wrong, whereas the other seems consistently right, then I should trust the latter. Furthermore, if we're in agreement about all the drawings I should judge her to be an epistemic peer. In this example,

my judgement of reliability is not made from an epistemically more secure position. I am in the same position as the two art historians. None of us had access to instruments that would enhance our reliability. I am judging their reliability by examining how well they match up with what I take to be the case.

Sometimes I may have good reason to assume that a person is reliable even though I don't have the opportunity to examine her track record. In such cases background theories I'm committed to may require me to judge that you are as reliable as I am or more reliable than I am in a given domain. For instance, given facts about the physical structure of your eyes and facts about how the human eye works, and the outside conditions necessary for good vision, I may be able to conclude that your vision reports ought to be reliable without any reference to these reports. Coupled with facts about my eyes, I can find out that we should be equally reliable.

Background theory based judgements of epistemic peerhood vary in strength and may sometimes be quite weak. Going back to our arithmetic example, knowing about your intelligence level and your educational attainments I can judge that you should be reliable in doing mathematical operations in your head. But this is a much weaker reason than the one your track-record would give me. The support relation between this background theory and the judgement that you ought to be reliable in doing mathematical operations in your head is also much weaker than the support relation between the background theory and the reliability judgement in the visual

reports example. We would be very surprised if two people an optician declared equally well-sighted were not equally reliable in their visual reports.[28] However, we wouldn't be surprised if two people with the same intelligence level and educational attainments were not equally reliable at doing mathematical operations in their head.

I should at this point remark on why I've defined an epistemic peer as I did. As you will recall, I defined an epistemic peer as someone whom I consider to be as reliable as I am in a certain domain. There are other definitions in the literature. For instance, according to Kelly

> 'two individuals are epistemic peers with respect to some question if and only if they satisfy the following two conditions: (i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and (ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias.'[29]

That two individuals satisfy the conditions set out by Kelly is a reason to expect that they will be equally reliable. Nevertheless, what matters is that they be equally reliable in the way I defined above. If the conditions laid out by Kelly did not contribute to one's likelihood of having true beliefs, why would we care about them? Alternatively, if someone can be reliable even though she does not satisfy Kelly's conditions, why should we not treat them as an epistemic peer? Suppose there is a person whom you find to

[28]Here I'm assuming that the optician reaches this conclusion by just examining your eyes and does not rely on your visual reports as they usually do.

[29]Kelly, 'Epistemic Significance', p. 174-5.

be thoughtless, and careless in her reasoning. However, when, for instance, doing sums in her head she gets the right answer as often as you do. That person should count as an epistemic peer in that domain, and her answers should have a role in your reasoning even though she doesn't satisfy Kelly's conditions for epistemic peerhood. She is a good indicator of what is the case in that domain. Kelly's, and other similar definitions, should be interpreted as background theories which would give us a reason to expect two people to be epistemic peers rather than as a definition of epistemic peer.

3.7 Judgement of Reliability and Response to Disagreement

After this, by no means exhaustive, survey of the ways in which we come to judge others as reliable, we can re-examine the Independence principle.

In cases where I've inferred your reliability from your track record by comparing what you take to be the case with what I take to be the case without occupying a more secure epistemic position, I should be entitled to violate Independence. For suppose that it would be wrong in those cases, where I've made track record judgements from the ground level, to violate Independence. Then, the order in which you and I expressed our judgements would be significant. But this is unacceptable.

To see this, suppose that we both express our judgement on a set of questions: $q_1, q_2, \ldots q_i \ldots q_n$. Let's say that our answers to $q_1 \ldots q_i$ are in agree-

ment. By looking at our track records I should consider you an epistemic peer. However, our answers to questions $q_i \ldots q_n$ are not in agreement. If I'm not allowed to violate Independence after I've decided that you are an epistemic peer, then I'm required to change my mind about my answers to $q_i \ldots q_n$ since an epistemic peer disagreed with my answers. Now, suppose that we began the sequence of questions from q_n instead of q_1. In this case, I wouldn't have identified you as an epistemic peer, because we disagreed on many questions. Accordingly, I would not be required to change my mind about my answers to $q_i \ldots q_n$. The order in which we begin the sequence of questions should not make a difference to what it is rational for me to believe.

Let us go back to our art historians example to illustrate this problem in a case. Suppose that I judged one of the art historians to be as reliable as I am. Then one day I notice that there was another set of Michaelangelo drawings I had access to, and this art historian has made judgements on their authenticity. I find that all of our judgements are in disagreement. If there's no independent reason for me to believe that his judgements on these occasions were less reliable, I'm required to change my mind about these drawings in order not to violate Independence. However, if, in the past I hadn't decided that this art historian was as reliable as I am, I could use my opinions about this new set of drawings to judge that this art historian was not reliable.

It should be clear from these examples that, insofar as we are justified in relying on track record judgements from the ground level to determine that another person is an epistemic peer, and this is the only basis of our judgement that we are epistemic peers, we should be entitled to violate Independence.

One possible response to this argument is to hold that it goes wrong in supposing that we can make a judgement regarding the other person's reliability before all the evidence is in. So we should make a judgement only when we've compared judgements on $q_i \ldots q_n$. The trouble with this response is that it will rob the Independence principle and the conciliatory response of all its teeth. For all instances of agreement and disagreement are relevant as evidence. If we are required to wait for all the evidence to come in before we judge that another person is an epistemic peer, when are we ever going to put Independence to use?

Are we entitled to rely on our judgements to determine whether someone is reliable in the way that we did in the art historians example? The natural answer is that we have no choice but to rely on our judgements.[30] Who else's judgements should we rely on? And if we are to rely on someone else's judgements, how are we to decide whose judgements we should rely on? We are inevitably led back to our own judgements as the final arbiter.

[30]David Enoch also makes this point in his paper, which presents an approach similar to mine. David Enoch, *Not Just a Truthometer: Taking Oneself Seriously (but not too seriously) in Cases of Peer Disagreement*, Unpublished manuscript ⟨URL: http://law.huji.ac.il/upload/truthometer.pdf⟩

I think that this is the right response to the question whether we are entitled to rely on our judgements in deciding whether someone is reliable. However, there is the need for a qualification. In certain cases, there have to be intermediary steps before the appeal to our judgements. Otherwise we would be epistemically irresponsible. Consider the test I suggested for determining whether someone is reliable at doing mathematical operations in their head. I suggested that we could do the operations on paper and check our results twice. Given that this method was available, it would be epistemically irresponsible of me to test your reliability by comparing the results you obtained when you did the mathematical operations in your head with the results I got when I did them in my head. This, however, does not mean that in the end I did not rely on my judgements. I relied on my judgement that doing mathematical operations on paper and checking the result was a reliable process. When there is the possibility of independent checks on our results and judgements we should rely on them, but ultimately some sources of knowledge admit of no independent checks.[31] In those cases we are entitled to rely on our judgements, which we make without taking up an epistemically more secure position, to determine that another person is an epistemic peer.

So, I conclude that we are entitled to rely on track record judgements from the ground level, and when, our judgement that the other person is an epistemic peer is based on such a method, we should be allowed to violate

[31]Van Cleve makes a similar point in a different context. See James van Cleve, 'Is Knowledge Easy - Or Impossible? Externalism as the Only Alternative to Skepticism', in: Steven Luper, editor, *The Skeptics: Contemporary Essays*, (Aldershot: Ashgate, 2003), p. 57.

Independence. If judgements from the ground level can be used to determine that someone else is an epistemic peer, they can also be used to undermine this initial judgement.

That we be should be allowed to violate Independence in these cases does not entail that we should always violate Independence. We leave some margin for error and invest different degrees of confidence in our judgements even when our judgements admit of no independent checks. So, if I identify another person as an epistemic peer without taking up an epistemically more secure position, and we disagree on something which I'm not very confident of, I should be reluctant to demote them from the status of epistemic peer. Similarly, if our past agreements were numerous and significant, I should be reluctant to demote you from the status of epistemic peer. The denial of Independence does not require us to deny these points. What the denial of Independence entails is that the disagreement itself can act as evidence in two ways, even when we lack independent grounds: (a) disagreement may give us reason to think that we may be wrong; or (b) disagreement can give us reason to think that the other person is not an epistemic peer.

Christensen examines a case that may be used to show that Independence can avoid the consequences we attributed to it in the case of track record judgements from the ground level.[32]

> Seminar: I'm in a meteorology graduate seminar with Stranger, another graduate student. I don't know him very well, but his

[32]I'm grateful to David Christensen for impressing on me the need to address this case.

> first few comments seem quite sensible to me. I take myself to be a pretty reliable thinker in meteorology, though not more reliable than most grad students. At the break, I discover that we've both read a fair amount about issue P, but while I'm quite confident that P, Stranger expresses equal confidence that $\neg P$. I, being a good Conciliationist, then become significantly less confident in P. But as the conversation develops, I find that Stranger and I disagree equally sharply about Q, R, S, T and so on–a huge list of claims. And these claims are not part of some tightly interconnected set of claims that would be expected to stand or fall together: they're largely independent of one another.[33]

Christensen concedes that the intuitively plausible response would be 'to stop putting so much stock in Stranger's claims'. However, he holds that if we fill out the rest of the case in a realistic way, we find that we get the right answer without violating Independence.

> Seminar, Cont'd: In addition to believing antecedently that I'm quite reliable in meteorology, and that the vast majority of others are about equally reliable, I also believe that there are a very few people–call them meteorologically deranged–who are horribly unreliable. I'm extremely confident that I'm not one of them. And I take such people to be rare enough that my estimation of my own reliability is not much different from my estimation of the reliability of a random person in the field.[34]

[33] Christensen, 'Question-Begging', p. 18.
[34] Christensen, 'Question-Begging', p. 19.

Once the case is filled out in this way, Christensen suggests we should reason as follows. Given the number of disagreements between us it's highly unlikely that both of us are reliable, so one of us has to be 'deranged'. And 'Since I'm more confident (independent of the disagreement) that I'm not deranged than that Stranger isn't, I should become more confident that I'm better at evaluating the evidence than Stranger is.'[35]

Christensen is correct to claim that this reasoning does not violate Independence. However, this reasoning is available because we have assumed that I have independent reason to believe that I'm more reliable. Christen's set up of the case involves the assumption that he has independent grounds to believe that he's not deranged. It is because Christensen can help himself to this assumption that he does not violate Independence. However, in cases where our reliance on track record judgements from the ground-level are justified, we won't have independent grounds to believe that we are reliable. Recall my claim on page 66 that we should rely on track record judgements from the ground-level only when we can't take up a more epistemically secure position. Reliance on track record judgements from the ground level is justified only when there are no independent grounds to establish reliability. As a result, the reasoning Christensen offers that respects Independence will not be available to us when our reliance on track record judgements from the ground-level is justified. In those cases the counter-intuitive results in Christensen's Seminar case and the second part of our art historians case can only be avoided by violating Independence.

[35]Christensen, 'Question-Begging', p. 19-20.

We can state the conditions for when reliance on track record judgements is justified and why the kind of reasoning Christensen offers in the Seminar case will be unavailable better by borrowing Alston's terminology of 'basicness'. According to Alston, a source of belief, *O*, is 'basic' iff 'Any (otherwise) cogent argument for the reliability of O will use premises drawn from O'.[36] In other words, when a belief source is basic its reliability can't be established by other belief sources. Now, my suggestion is that when a source of belief is basic, we can come to judge that another person's judgements based on that same source of belief are reliable only by relying on track record judgements from the ground level. The fact that this source of belief in question is basic entails that we will lack independent grounds for judgements of reliability of the kind that figured in Christensen's Seminar case.[37]

Let us turn to other modes of judging another person to be an epistemic peer. Should we respect Independence when our reliability judgement is based on background theories? The key point to keep in mind, throughout, is that we have no choice but to rely on our judgements and that our judgement that another person is an epistemic peer is fallible like our other judgements. The intuitive appeal of Independence lies in the very good question 'What's so special about you?' No doubt, this question has force. However, from the first person-perspective there's another question, which also has force, that also needs to be asked: 'What's so special about my relia-

[36]William P. Alston, 'Epistemic Circularity', *Philosophy and Phenomenological Research*, 47 September (1986):1, p. 8.

[37]Note that a non-basic source of belief may be circumstantially basic. Even though there may be arguments that can establish its reliability without employing its own deliverances, those resources may be unavailable under certain conditions. Track-record judgements from the ground level are justified in such cases too.

bility judgement?' The answer to our question whether Independence needs to be respected in the case of reliability judgements based on background theories, therefore, depends on our confidence in the background theory and how strongly our background theory supports the judgement that the other person is an epistemic peer. When we are confident of our background theory, and of the judgement that it gives us reason to identify the other person as an epistemic peer, we have good reason to respect Independence. This point is well-illustrated by the Horse Race case. In that case we have good reason to think that our judgement may be mistaken, and our background theory gives us good reason to believe that the other person is an epistemic peer. When, however, the relationship of support between our background theory and our judgement of epistemic peerhood is not as strong, or when we are not very confident of our background theory, we don't need to respect Independence.

In cases where we have background theories supporting our judgement of reliability, Independence draws support from the apparent epistemic symmetry between us and our epistemic peer. The thought is something like the following. My epistemic peer and I satisfy certain conditions, C. C is the main determinant of one's likelihood of arriving at true beliefs. Therefore, I should expect my epistemic peer to be as likely as I am to have true beliefs. Independence, quite plausibly, suggests that when we are confident of the fact that we both satisfy C, and of the fact that C is the main determinant of one's reliability, then we shouldn't rely on the belief under contention to demote the person we disagree with from the status of epistemic peer. How-

ever, when we are more confident of the judgement under contention than these premises, then violating Independence is more reasonable, because our belief in the existence of epistemic symmetry between us is weaker.

Should we be allowed to violate Independence when our knowledge of another's status as an epistemic peer is based on our evaluation of their track record from a more secure epistemic position? In such cases, the previous argument, which I gave to show that we should be allowed to violate Independence when our track record judgements are from the ground level, does not apply. Furthermore, in the case of track record judgements from a more secure epistemic position, our judgement that p would be less reliable than our judgement that we are epistemic peers. Accordingly, we should respect Independence in these cases.

I'd like to make two general points before turning to disagreements with many potential epistemic peers. The position I've defended that allows me to demote others from the status of epistemic peer because they disagree with me may seem epistemically arrogant. The following thought, however, should mitigate the appearance of epistemic arrogance. When I decide that you were not, after all, an epistemic peer my reasoning is not that I'm a better epistemic agent, so I'm entitled to discount your opinion. Rather, I think that my judgement that we are epistemic peers is more likely to be wrong than my judgement that p. Note also that I don't need to assume that my judgement that p cannot be mistaken. It is just that my belief that you're an epistemic peer is more likely to be mistaken. Even though this line

of reasoning displays epistemic arrogance in one respect, it shows epistemic modesty in another.

When I demote another person from the status of epistemic peerhood, is it not possible that I am making a mistake and foregoing a chance to improve my beliefs? There is, of course, this possibility. But, there is also the possibility of mistakenly treating another as an epistemic peer. The best that we can do is to be guided by what we take to be the best reasons we have. Judging that another person is not an epistemic peer is no less risky than judging that another person is an epistemic peer.

3.8 Disagreeing with many

One question we need to consider is the relevance of numbers to my argument.[38] Let us begin with background theory based judgements of reliability. In the previous section, I suggested that if we were confident in our background theory, and the relationship of support between the background theory and the judgement of epistemic peerhood was strong, we had good reason to respect Independence. And, if these conditions didn't hold, we were entitled to violate Independence. Here's one line of reasoning which suggests that numbers ought to make a difference.

According to *Condorcet's jury theorem*, if we assume that (a) every voter has a more than even chance of arriving at the correct answer on a given question, and (b) the probability that any given voter arrives at the cor-

[38]I'm grateful to David Miller for impressing on me the need to address this question.

rect answer is independent of the probability of the other voters arriving at the correct answer; then the probability that the answer voted for by the majority is correct converges to 1 as the number of voters approaches infinity.[39] Condorcet's jury theorem offers a much more general reason to take the opinions of others into account than epistemic peerhood. The insight of the theorem is that numbers can take over the role of expertise. All that is needed is each voter to have a better than even chance. Condorcet's jury theorem can take some of the weight our background theory and the relationship of support between it and the judgement of reliability have to bear. All that the background theory needs to establish is that the many individuals we are agreeing with have a reliability that falls within a certain range.

There are several considerations which bear on the question of whether it is rational to stick to one's opinion in the face of disagreement with the majority. In the case of background judgements of reliability, I won't focus on Independence. I'll be focusing on the question of whether it is a good epistemic policy to defer to the majority, and try to raise some general doubts about a policy of deferring to group judgements.

One central consideration is the nature of the claim that is being put the vote. Suppose the claim under consideration is 'Junior academic A should be granted tenure', C. Everyone agrees that A should be granted tenure if and only if she is excellent in research, p, and she is excellent in teaching, q, and

[39]For an accessible presentation of the theorem see David M. Estlund, 'Opinion leaders, independence, and Condorcet's Jury Theorem', *Theory and Decision*, 36 March (1994):2.

she has contributed to the department's life, r.[40] Suppose that (p & q & r) is true.[41] Let's say the probability of every individual of getting each of the propositions right is 0.51. Then, the probability that they will get all of the propositions right is $0.51^3 \approx 0.133$.[42] In this case, the probability that the majority reaches the correct verdict on whether C or not-C approaches 0 as the size of the public increases. It is only if the probability of every individual getting each of the propositions right is more than $0.5^{1/3} \approx 0.79$ that the probability of their getting the truth value of the conjunction right will be more than 0.5, which is what's needed for the application of Condorcet's jury theorem. Now, C is not a very complex claim. Many of the the questions we are likely to disagree with others on will be more complex. So, the demand on individual reliability will increase.

We could pursue a different method of aggregating the judgements of the department on the question of whether C or not-C. Instead of asking individuals their beliefs about the conjunction, we could ask them about the conjuncts. In terms of its ability to arrive at the correct answer, this proce-

[40]This example is an adaptation of an example offered by List in Christian List, 'The Discursive Dilemma and Public Reason', *Ethics*, 116 (2006):2.

[41]I need to make this assumption, because the probability of the majority arriving at the correct answer when (p & q & r) is the case is not the same as the probability they will arrive at the correct answer when ¬(p & q & r) is the case. In the latter case, they may get the result while being mistaken about the conjuncts. See Philip Pettit and Wlodek Rabinowicz, 'The Jury Theorem and the Discursive Dilemma.' *Philosophical Issues (supplement to Nous)*, 35 (2001), p. 296.

[42]Here and for the rest of this discussion I'm making two simplifying assumptions. Firstly, I'm assuming that each of the premises are independent. For instance, learning that p is true does not tell us anything about whether q is true or not. Secondly, I'm assuming that the probability of an individual getting any given premise right is independent of the probability of them getting the other premises right. For instance, the probability that a committee member arrives at the correct answer on whether p or not-p is independent of the probability that she arrives at the correct answer on whether q or not-q.

dure is far superior to asking others for their judgements on the conjunction. Again suppose that (p & q & r) is true, and the probability of every individual getting each of the propositions right is 0.51. In this instance, Condorcet's jury theorem will be applicable to each of the conjuncts. As numbers increase, the probability that the majority got the correct truth value for p, q, and r individually will approach 1. As we saw in the previous paragraph, the probability that the majority got the correct answer for whether C or not-C approached 0 under the same assumptions. This will generate a contradiction. There will be a majority in favor of each of the following: p, q, r, and ¬ (p & q & r). To see intuitively how this can come about consider the following table:

Table 3.1: *The Discursive Dilemma*

	p	q	r	C (p & q & r)
Member 1	F	T	T	F
Member 2	T	F	T	F
Member 3	T	T	F	F
Majority	T	T	T	F

Here, we have a majority in favour of p, q, r, and ¬ (p & q & r). There is a majority in favour of an inconsistent set of propositions. This is an instance of what Pettit has called the *discursive dilemma*.[43] The discursive dilemma shows that deferring to the majority on both p, q, r, and (p & q & r) is not

[43]Philip Pettit, 'Deliberative Democracy and the Discursive Dilemma', *Philosophical Issues (supplement to Nous)*, 11 (2001).

a good epistemic policy and will cause us to have inconsistent beliefs.[44]

One way to avoid this outcome is by deferring to the majority only on the premises, or only on the conclusion. In the literature, the procedure of aggregating the public's judgements by taking a majority vote on the conclusion has been called the *conclusion-based procedure*. The procedure of aggregating the public's judgements by taking a majority vote on each of the premises, and arriving at the conclusion based on the results of these results, has been called the *premise-based procedure*.[45] As we have seen the premise-based procedure is better at tracking the truth.[46] When we followed the premise-based procedure the probability that the majority arrived at the right answer approached 1 as the numbers increased. When we, under the same assumptions, followed the conclusion-based procedure the probability that the majority arrived at the right answer approached 0. The possibility of the discursive dilemma only establishes that we shouldn't defer to the majority's views on both the conclusion and the premises. The failure of the conclusion-based procedure at tracking the truth counts against that procedure, but the premise-based approach seemed to do well. This leaves the premise-based procedure as a plausible candidate.

[44]Pettit also argues this. See Philip Pettit, 'When to defer to majority testimony - and when not', *Analysis*, 66 July (2006):3.

[45]List, 'The Discursive Dilemma'.

[46]This statement actually needs to be qualified. There can be cases where the conclusion-based procedure is more likely to produce the correct answer. However, that happens because the conclusion-based procedure can produce the right answer for the wrong reasons. For a detailed discussion see Luc Bovens and Wlodek Rabinowicz, 'Democratic Answers to Complex Questions - An Epistemic Perspective', *Synthese*, 150 May (2006):1.

In light of this, deferring to the majority by employing the premise-based procedure might seem to be a good epistemic policy. However, the premises of the initial argument themselves can be the conclusions of another argument. In that case we can run the same argument we ran against the conclusion-based procedure in the case of C to the premises. Let's say the following is true of research excellence: someone's research is excellent if and only if it shows originality, s, and has had impact in the field, t, that is, p $\leftrightarrow$ (s & t). For the majority to arrive at the correct truth value for (s & t) on the conclusion-based procedure, they have to have a more than $\sqrt{0.5}$ chance of getting the truth value of each of the conjuncts right. It may be suggested that we can rely on the premise-based procedure again to arrive at the majority's views on *s* and *t*. But *s* and *t* can also be the conclusions of other arguments in which case the same problem will reappear.

I think our discussion of this example establishes two main points. Firstly, deferring to the majority only on the premises in order to avoid ending up with inconsistent beliefs is not a feasible strategy. The premises themselves will often be conclusions of other arguments, and the premises of those arguments will often be the conclusion of other arguments, and so on. An immediate point which follows from this is that in the case of many of our beliefs it is not possible to keep revisions local and preserve consistency. A given belief will figure in the argument for several other beliefs, and those beliefs in their turn will figure in the argument for others. Furthermore, all of these beliefs will be open to challenges by the majority, making the policy of deferring to the majority intractable. This response goes a long way

towards explaining why deferring to the majority is a bad epistemic policy. However, it does not wholly assuage the worry that Condorcet's theory gives rise to. That worry is assuaged I believe by the second point established by this example. In order to establish that the majority is likely to arrive at the correct answer on a given question, one needs to make several complex assumptions about their reliability, and at some point assume that they are likely to be very reliable. For instance, in the previous example, we needed only to assume that the public had a more than even chance of arriving at the correct truth values for p, q, and r, because we employed the premise-based procedure. However, if we employ the conclusion-based procedure for (s & t) which gave us the truth value of p, then we need to assume that the probability that they arrive at the correct answer on s is greater than $\sqrt{0.5}$ and the probability that they arrive at the correct answer on t is greater than $\sqrt{0.5}$.[47] For Condorcet's jury theorem to have force for us, we need to make very demanding assumptions about others' reliability. In addition to this, we need to assume that we are reliable at judging the reliability of others, and our ability to ascertain that they have arrived at their judgements independently. This information will actually be very difficult to come by, and compared to the information we can usually obtain about the premises themselves, unreliable. Consider the committee case, and in particular the

[47] As an illustration of how reliable we need to assume others to be when the propositions become more complex, suppose that, in addition to s and t, there were five more factors which were necessary for research excellence. Suppose that we are going to rely on the conclusion-based procedure. We will ask everyone what they believe is the case regarding p, rather than asking them what they believe is the case on s, t etc. Assuming that p is true, for the majority to arrive at the correct answer through this procedure, everyone would have to have a greater than $0.5^{1/7} \approx 0.91$ chance of arriving at the correct answer on each of the premises.

judgement about research excellence, again. I can know about myself how thoroughly I've studied the candidate's research, and whether I'm in a position to evaluate whether the candidate's work is original. However, I can't know this about 10 other committee members.

Let us, finally, turn to cases where our judgement that others are reliable is based on track-record judgements from the ground level and Independence. Suppose there is more than one epistemic peer I disagree with, and I've judged that they are epistemic peers by relying on track record judgements from the ground level and I didn't have any antecedent reason to think that they were reliable. Am I still entitled to violate Independence? The argument for violating Independence given in 2.7 applies to the individual cases, and so should apply cumulatively to the disagreement with many epistemic peers. If I'm entitled to rely on track-record judgements from the ground level to determine that others are reliable, that is, if judgements from the ground level can count towards establishing several people's reliability, then they should also count towards establishing their unreliability.

This result initially seems counter-intuitive. However, once we carefully distinguish between different possibilities and set out what our entitlement to violate Independence entails, we can remedy this. Consider the following two cases. In the first case, I've compared judgements with several people on a few cases. I then find that there are several new cases where we disagree. In the second case I've compared judgements with several people on many cases. I find that in a new case we disagree. In the first case, the initial

instances of agreement give me reason to judge them to be epistemic peers. However, the many cases of disagreement give me reason to doubt my initial judgement that they were my epistemic peers. Assuming that I had judged these people to be epistemic peers only given the initial run of agreements, I can change my mind about their status as epistemic peers only by violating Independence. Given this set up of the case, demoting them from the status of epistemic peers seems like the intuitively correct response. In the second case, the many instances of agreement give me good reason to judge them to be epistemic peers. Against this background, it will be wrong to demote them from the status of epistemic peer given a single case. However, this is in line with our treatment of the single person case. Even in the single person case, when there are many significant past agreements, we should be reluctant to demote the other person when there are few new disagreements. The fact that we're entitled to violate Independence says something about evidence that's admissible in our reasoning. It does not say anything about the role that evidence should play in all the instances that this evidence is available. In certain cases, a person's disagreement with us can give us a reason to demote them from the status of epistemic peerhood. In others, the disagreement may give us reason to change our minds.

3.9 Conclusion

What are the implications of our account of when we are required to respect Independence and when we are entitled to violate it to disagreements in moral and political philosophy? When someone has different intuitions than us, can

we justifiably remain as confident as before without, as Sidgwick required, 'prov[ing] independently that the conflicting intuitor has an inferior faculty of envisaging truth in general or this kind of truth'? If our previous account is correct, the answer to this question depends on how we determine the reliability of others in the domain of morality.

I would suggest that our determination of the reliability of others will be based mostly on track-record judgements from the ground level. This is because the other two methods of judging reliability are unavailable. We are very far from having a theory of philosophical and moral cognition that would give us a reason to confidently judge that another person is an epistemic peer. We also don't have a more secure epistemic position to judge our and others' reliability on moral questions than what we believe on due reflection.

To see the shortcomings of our background theories compare the conditions in Kelly's definition of epistemic peer I quoted on page 62, which is pretty much all we can rely on to identify another as an epistemic peer in moral philosophy, and what we can say about two people being equally reliable in their visual reports. Not only is this a less developed account than the account we would have in other domains, it is also very difficult to determine that two people satisfy these conditions. Take Kelly's first condition that epistemic peers 'are equals with respect to their familiarity with the evidence and arguments which bear on that question'. How confidently can we know that two people fulfil this condition with respect to a given philosophical question?[48]

[48] A further difficulty is that which arguments are relevant will often be a point of contention.

(In addition to the fact that we don't have a satisfactory theory, it's also unlikely that we can have a theory that's free of ground level judgements. The difficulty of coming up with an account of competent moral judges that doesn't involve moral qualities and judgements illustrates this difficulty.)

Two results follow from the fact that our judgements of the reliability of others in the domain of morality will be based on track-record judgements from the ground level. Firstly, the fact that someone doesn't share our intuitions, by itself, will not be a reason for us to lose our confidence in those intuitions. Disagreement with others matters only against a background of past agreements. Secondly, in cases where we disagree with someone we've judged to be an epistemic peer, we are entitled to violate Independence. If we go back to the congruence requirement, we see that Sidgwick's view can support only a significantly restricted version of it. Our epistemic peers are people with whom we agree on a wide range of cases, and when we disagree with them sometimes it will be more reasonable to revise our initial judgement of epistemic peerhood rather than the judgement under contention.

In section 3.3 I argued that, in the absence of the congruence requirement, what we could justify to others and what we are rationally entitled to believe could differ significantly. In the rest of this chapter, we looked at the most plausible way to support the congruence requirement. We have seen that this defence based on the epistemic significance of disagreement supports the congruence requirement in a very limited range of cases. The upshot of all this is that in most instances of disagreement about morality what we

are rationally entitled to believe and what we can justify to others differ significantly. In the next chapter, I will argue that this gap explains some of the specific features of Rawls's constructivism in his later works, and that we have reason to expect the theory of justice that we ought to accept to differ from the set of principles that can be justified to the adherents of different reasonable comprehensive doctrines.

Chapter 4

Constructivism

4.1 Introduction

The first aim of this chapter is to offer a characterization of constructivism and to delineate some of the main ways in which constructivist theories vary. My characterization of constructivism will inevitably be contentious, because it is a difficult doctrine to pin down. One reason for this difficulty is that it is a new doctrine, and its proponents are still in the process of giving it shape. A second reason is that it is offered as an alternative to doctrines which are themselves hard to pin down. For instance, construed as a metaethical doctrine, constructivism is often presented as an alternative to realism. However, how best to understand realism about ethics is a difficult question with no agreed upon answer. These difficulties carry over to accounts of constructivism. I believe my characterization covers the cases that are central for my purposes and brings forth the relevant differences between constructivist doctrines. My second aim, in this chapter, will be to apply the accounts of

the previous chapters to Rawls's constructivism(s). I shall argue that Rawls's constructivism in *A Theory of Justice* and the form of constructivism in his later works differ, because the first project aims to offer an account of justice that is acceptable from the perspective of the individual inquirer and the second project aims at offering a set of principles that will be acceptable to the adherents of different reasonable comprehensive doctrines. I shall argue that the tensions we identified in the previous chapter between what we can justify to others and what we are rationally entitled to believe carry over to these two constructivist projects. I shall also argue that it is the argumentative strategies Rawls adopts to arrive at a set of principles that will be acceptable to the adherents of different reasonable comprehensive doctrines rather than constructivism as such that makes Rawls's theory vulnerable to Cohen's charge that it 'generates misteachings about justice and other values'.[1]

4.2 Bound vs Unbound Constructivism

Some philosophers think that the basic normative notion is reasons. Other normative concepts are to be analysed in terms of reasons. Others think that goodness, or value, is the basic normative notion. What I will call *bound constructivism* offers a reduction of reasons, or values, within a given domain in terms of other reasons, or values. Bound constructivisms take the existence of reasons or values in some domains as given and seek to offer an account of reasons or values in other domains based on these. For

[1]G. A Cohen, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008), p. 300.

this reason, they do not constitute metaethical positions. One can give a non-constructivist account of the reasons which make up the reduction base. *Unbound constructivism* on the other hand proposes to offer an account of what it sees as the basic normative notion whether it is reason, goodness, or value *simpliciter*.[2]

Most common forms of bound constructivism are captured by the following formula, which identifies the reasons of agents in a certain domain with the reasons of other agents, who are under specific conditions, in another domain. (For ease of exposition I shall focus only on reasons for the rest of this chapter.):

> R is a reason of domain D_1 for subject(s) S_1 if and only if R is a reason of domain D_2 for subject(s) S_2 under conditions C.

Often, principles play a mediating role between reasons in different domains. In those cases we can use the following formula:

> R is a reason of domain D_1 for subject(s) S_1 if and only if R follows from principles subject(s) S_2 under conditions C would have reasons of domain D_2 to choose.[3]

[2]Several other recent works on constructivism draw a distinction similar to the one I draw between bound constructivism and unbound constructivism. See Sharon Street, 'Constructivism about Reasons', in: Russ Shafer-Landau, editor, *Oxford Studies in Metaethics*, Volume 3, (Oxford: Oxford University Press, 2008); David Enoch, 'Can there be a global, interesting, coherent constructivism about practical reason?' *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 12 (2009):3.

[3]This formula can be varied to account for some variations in constructivist theories. Specifically, 'reasons of domain D_2 to choose' can be replaced by 'reasons of domain D_2 not to reject' to accommodate Scanlon's theory.

Gauthier's constructivism that offers a reduction of moral reasons into reasons of self-interest, suitably specified, can be (roughly) formulated as follows:

> R is a reason of morality for rational persons who can cooperate with others on mutually beneficial terms if and only if R follows from principles which fully rational subjects would have prudential reason to agree to under conditions necessary for voluntary agreement.[4]

And Rawls's later constructivism can, preliminarily, be formulated as follows:

> R is a reason of justice for citizens of a democratic society if and only if it follows from principles that rational persons represented as free and equal moral persons would have conclusive prudential reason(s) to choose under fair conditions to govern the basic structure of their society.

Rawls's constructivism rests on a normative basis. Rawls takes the the idea of a well-ordered society as a fair system of cooperation between reasonable and rational citizens as a given. He acknowledges this when he writes:

> '... [N]ot everything... is constructed; we must have some material, as it were from, which to begin... The procedure itself is simply laid out using as starting points the basic conceptions of society and person, the principles of practical reason, and the

[4]David Gauthier, *Morals by Agreement*, (Oxford: Clarendon Press, 1986).

public role of a conception of justice.'[5]

He also explicitly mentions that the choice in the original position is based on reasons and is not an instance of 'radical choice', that is, 'a choice not based on reasons'.[6] Rawls also takes it as given that the addressees of his constructivist procedure already have moral commitments, and that they take themselves to be free and equal. So not only does Rawls assumes the existence of reasons that appear on the right hand side of the bound-constructivist formula, he also assumes that the addressees of his constructivist formula are committed to the existence of reasons which underlie the constructivist formula. His addressees take themselves to be free and equal, and have other moral reasons.

Unbound constructivism, if it is going to be an independent metaethical doctrine, needs to avoid statements about reasons which appeared in the right-hand side of the bound constructivist formula. It is here that things become particularly contentious. Here is one attempt:

> R is a reason *simpliciter* for subject(s) S_1 if and only subject(s) S_2 under conditions C would have attitude A toward it.

The challenge for the proponent of unbound constructivist here would be to provide a characterization of the right hand side of the above formula which ensures that (a) it does not collapse into bound constructivism; and (b) it does not collapse into naturalism.

[5]John Rawls, *Political Liberalism*, New edition. (New York: Columbia University Press, 1996), p. 104.

[6]John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 354.

I shall not pursue the discussion of unbound constructivism any further. I am primarily interested in bound constructivism. Unbound constructivism is not a rival to bound constructivism as long as it does not make first-order claims which are in conflict with the claims of bound constructivism. This is because bound constructivism is not a metaethical view. It is not committed to a metaethical account of the reasons which function as the reduction-base.[7] In cases where unbound constructivism makes first-order claims which conflict with the claims of a successful bound-constructivist theory, it faces an uphill struggle. Since the bound constructivist theory is assumed to be successful as a first-order account, the unbound constructivist theory will have to rely on metaethical claims to outweigh the first-order claims it conflicts with. As I argued in section 2.4, it is unlikely that one can find metaethical claims which can outweigh ethical claims.

As a first-order account, bound constructivism is not motivated by epistemological and ontological worries. Its motivation is to provide a unitary and illuminating account of the moral judgements in the target domain, and show how these judgements flow from more basic judgements. Different kinds of metaethical accounts can be given of the reasons in the reduction base. For instance, we can adopt a realist, unbound constructivist, etc. account of Rawls's assumption that people ought to be treated as free and equal.

[7]Bound constructivism, of course, takes the reasons which function as the reduction base to be valid. However, it doesn't provide an account of their validity.

4.3 Radical vs Conservative Constructivism

Some possible ways in which unbound constructivist theories vary is clear from the formulation of the general unbound constructivist formula. Constructivist procedures vary in the domain they seek to reduce, their addressees, the way the agents of construction are characterized, the conditions the agents of construction are under, and the kind of reasons the agents of construction have.

Some other variations, are not obvious from this formulation. Constructivist theories differ on their treatment of the judgements in the domain they offer accounts of. Some versions of constructivism, which I shall call *conservative constructivism*, seek to conserve most of our judgements in the domain they offer an account of, while allowing room for some revisions. Rawls's constructivism is an example of this approach. The principles Rawls's constructivist procedure generates are tested against our considered convictions about justice.[8]

Other versions of constructivism, which I shall call *radical constructivism*, do not seek to conserve our judgements in the domain they offer accounts of. For instance, according to Gauthier's theory those who cannot cooperate, such as 'the unborn, the congenitally handicapped and defective', 'fall beyond the pale of morality'.[9] If our intuitions fail to match Gauthier's theory, it is our intuitions which have to give way, because they have no weight 'if

[8]There's of course room for revision at both ends.
[9]Gauthier, *Morals by Agreement*, p. 268.

morality is to fit within the domain of rational choice'.[10] Nevertheless, the constructivist who does not seek to conserve our intuitive judgements cannot jettison all our judgements. The constructivist about a given domain needs to show that she is offering a constructivist theory of reasons in that domain. For the constructivist to be able to show this some of our existing beliefs about that domain need to be preserved. So, Gauthier needs to preserve some of our beliefs about morality in order to show that he is providing a constructivist theory about morality rather than something else.

The metaphor of construction, and some of Rawls's comments in 'Kantian Constructivism in Moral Theory', might suggest that Rawls did not always have the goal of preserving our judgements about justice, and furthermore that the absence of this goal is a defining feature of constructivism. For instance, Parfit writes:

> 'According to *intuitionists*, Rawls writes, there are certain independent truths about which acts are wrong, and about which facts give us reasons. Two examples are the truths that slavery is wrong, and that we have reasons to prevent or relieve suffering. These truths are *independent* in the sense that they are not created or constructed by us. According to a different view, which Rawls calls *constructivism*, there are no such truths. On this view, what is right or wrong depends entirely on which principles it would be rational for us to choose in some Kantian or contractualist thought-experiment. In Rawls's phrase, it's for us

[10] Gauthier, *Morals by Agreement*, p. 269.

to decide what the moral facts are to be. If we are constructivist contractualists, and we decide that it would be rational to choose principles that permit slavery, we ought to conclude that slavery is not wrong. Though slavery may seem to us to be wrong, constructivists reject appeals to our moral intuitions, which some of them claim to involve mere prejudice, or cultural conditioning, or to be mere illusions.'[11]

Parfit's comments apply to radical constructivism, but not to conservative constructivism. Radical constructivism's treatment of the judgements in the domain that it offers a reduction of is not definitive of constructivism. Conservative constructivisms put forward the hypothesis that once we arrive at the correct account of reasons within a given domain, we will come to see that the reasons in that domain are given by a constructivist procedure. Speaking metaphorically, we can say that these constructivists *discover* that reasons within a given domain are constructed from reasons from another domain.

Parfit reads Rawls as a radical constructivist, but Rawls is better understood as a conservative constructivist. In his discussions of constructivism Rawls makes it clear that the difference between constructivism as he understands it and rational intuitionism is not the role of our moral intuitions in these doctrines. He writes:

'... [B]oth constructivism and rational intuitionism rely on the

[11]Derek Parfit, *On What Matters*, ⟨URL: http://users.ox.ac.uk/~ball2568/parfit/parfit_-_on_what_matters.pdf⟩, p. 288, emphasis in the original.

idea of reflective equilibrium. Otherwise intuitionism could not bring its perceptions and intuitions to bear on each other and check its account of the order of moral values against our consider judgements on due reflection. Similarly, constructivism could not check the formulation of its procedure by seeing whether the conclusions reached match those judgements.'[12]

The difference between these constructivists and rational intuitionists lie not in their employment of intuitions, but in their account of the final result. They disagree on the order of explanation:

'Rational intuitionism says: the procedure is correct because following it correctly usually gives the correct (independently given) result. Constructivism says: the result is correct because it issues from the correct reasonable and rational procedure correctly followed.'[13]

Since conservative constructivism seeks to preserve our judgements in the domain that is being reduced, the failure to give a constructivist account of the domain that is the candidate for reduction is to be taken as a failure on the part of constructivism rather than as symptomatic of a problem with the domain whose reduction is attempted. Rawls accepts this when he writes:

'Of course, the repeated failure to formulate the procedure so that it yields acceptable conclusions may lead us to abandon political constructivism. It must eventually add up or be rejected.'[14]

[12]Rawls, *Political Liberalism*, p. 96.

[13]John Rawls, *Lectures on the History of Moral Philosophy*, (Cambridge, Mass: Harvard University Press, 2000), p. 242.

[14]Rawls, *Political Liberalism*, p. 96n.

In the case of constructivism about justice, conservative constructivism is preferable to radical constructivism, because we have beliefs about justice which we are very confident of. A theory about justice, if it is going to be acceptable to us, has to conserve these beliefs.

Conservative constructivism can be compared to the logicist project in the philosophy of mathematics.[15] The logicists' aim was to show that arithmetic could be reduced to pure logic. Since mathematics had been reduced to arithmetic this reduction would, in turn, entail the reduction of mathematics to logic. The logicists did not seek to reform existing mathematical practice. What they sought to do was show how mathematics as it existed could be shown to be derivable from logic. If successful, the logicist project would not have an impact on mathematics, but it would produce significant epistemological and metaphysical results. It would show how we can have knowledge of mathematics in virtue of our knowledge of logic, and it would show that we do not need to postulate the existence of mathematical objects. When the logicist project failed, the proponents of logicism did not conclude 'So much the worse for mathematics', but either gave up the logicist project or tried to amend the logicism.

We should note that the distinction between conservative and radical constructivism is distinct from the distinction between bound and unbound constructivism. Accordingly, we can have different combinations of these views.

[15]For a brief and helpful overview of the logicist project, which is the one I'm following here, see Scott Soames, *Philosophical Analysis in the Twentieth Century*, Volume 1, (Princeton, N.J: Princeton University Press, 2003), p. 132-164.

Rawls's constructivism exemplifies bound and conservative constructivism. This is the type of constructivism, I shall defend in the next chapter.

4.4 Constructivism in *Political Liberalism*

Both *Theory* and *Political Liberalism* rely on the original position as a constructivist procedure, which as we've seen contains moral notions. And in both works, Rawls claims to adhere to the method of reflective equilibrium. Accordingly, both works may seem to be examples of bound and conservative constructivism. In this section, I shall argue that the role of reflective equilibrium in the two works is different, and Rawls's approach in *Political Liberalism* can't be straightforwardly characterized as an instance of conservative constructivism. This fact, I shall argue, is to be explained by the fact that Rawls addresses different questions in the two works and employs different argumentative strategies required by the these two questions.

Whereas in *Theory* Rawls was concerned mainly with justice, *Political Liberalism* is mainly concerned with legitimacy.[16] This concern with legitimacy requires a significant change in perspective. The perspective of *Theory* is that of the individual inquirer asking himself what justice is. In *Theory* only the views of the reader and the author counted.[17] The perspective of *Politi-*

[16]This has been observed by several commentators. See, for instance, David Estlund, 'The Survival of Egalitarian Justice in John Rawls's Political Liberalism', *The Journal of Political Philosophy*, 4 (1996):1, p. 68; Allen Buchanan, 'Justice, Legitimacy, and Human Rights', in: Victoria Davion and Clark Wolf, editors, *The Idea of a Political Liberalism: Essays on Rawls*, (Lanham: Rowman and Littlefield, 2000), p. 73.

[17]John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 44.

cal Liberalism is the interpersonal perspective of citizens trying to arrive at principles they can justify to one another. The addressee of its arguments are the citizens of democratic societies. *Political Liberalism* aims to uncover 'a public basis of justification on questions of political justice'.[18]

On the focus on the individual inquirer in *Theory* Rawls writes:

> 'TJ [*Theory*] does not address persons as citizens but rather as individuals trying to work out their own conception of justice as it applies to the basic political and social institutions of democratic society. For the most part their task is *solitary* as they reflect on their own considered judgements with their fixed points and the several first principles and intermediate concepts and the ideals they affirm. TJ is presented as a work individuals might study in their attempt... to attain the self-understanding of wide reflective equilibrium.'[19]

The change in perspective is evident throughout *Political Liberalism*. Here are two typical passages as evidence:

> 'The aim of justice as fairness... is practical: it presents itself as a conception of justice that may be shared by citizens as a basis of a reasoned, informed, and willing political agreement. It expresses

[18]Rawls, *Political Liberalism*, p. 100.

[19]Rawls cited in Norman Daniels, *Justice and Justification: Reflective Equilibrium in Theory*, (Cambridge: Cambridge University Press, 1996), p. 145, emphasis added. This quotation is from an unpublished piece by Rawls. For those interested, the full citation details Daniels offers are as follows: John Rawls, '*A Theory of Justice* and *Political Liberalism* and Other Pieces: How Related?', Draft of comments presented at October 1995 Santa Clara Conference on Rawls's work. '

their shared and public political reason.'[20]

> '... [T]he aim of political liberalism is to uncover the conditions of the possibility of a reasonable public basis of justification on fundamental political questions. [21]

The shift of focus from justice to legitimacy brings about this change in perspective because of the liberal principle of legitimacy. According to the liberal principle of legitimacy,

> '... [O]ur exercise of political power is proper and hence justifiable only when it is exercised in accordance with a constitution, the essentials of which all citizens may reasonably be expected to endorse in light of principles and ideals acceptable to them as reasonable and rational'.[22]

In order to show that the principles he proposes pass this test, Rawls has to take into account the different viewpoints of reasonable persons. In *Political Liberalism* Rawls closely follows the model of interpersonal justification we looked at in 3.2. His strategy is to present justice as fairness as a free-standing political conception that does not rely on any comprehensive view so that it may be the subject of an overlapping consensus among members of various reasonable comprehensive doctrines. Rawls thinks that such a strategy ensures stability in the face of reasonable pluralism and ensures the legitimacy of a liberal state. Rawls's theory is not a comprehensive

[20]Rawls, *Political Liberalism*, p. 100.
[21]Rawls, *Political Liberalism*, p. xxi.
[22]Rawls, *Political Liberalism*, p. 217.

doctrine, because (a) it only applies to the basic structure of a society; (b) it confines itself to political questions; and (c) 'it is not formulated in terms of a general and comprehensive religious, philosophical, or moral doctrine but rather in terms of certain fundamental intuitive ideas viewed as latent in the public political culture of a democratic society'.[23] Rawls employs notions accepted by the relevant parties, the citizens of democratic societies, in his construction. The original position is claimed to systematize ideas, such as the idea of 'society as a fair system of cooperation over time, from one generation to the next', the idea of citizens as free and equal, and the idea of a well-ordered society, that are implicit in the public culture of democratic society.[24] The outcome of the original position is supposed to be acceptable to citizens, because it models ideas that they already affirm, and it is not dependent on any comprehensive doctrines.

Rawls's explicit strategy in *Political Liberalism* is a *strategy of exclusion*. Those beliefs which are matters of contention between members of reasonable comprehensive doctrines are excluded in the argument for and in the construction of Rawls's theory. The strategy of exclusion excludes not only comprehensive doctrines, but also certain non-philosophical beliefs. One example is the restriction on beliefs 'about human nature and the way political and social institutions generally work, and indeed all such beliefs relevant to political justice'.[25] Rawls requires that we ascribe to the denizens of the original position 'the procedures and conclusions of science and social thought,

[23]Rawls, *Collected Papers*, p. 423-7.
[24]Rawls, *Political Liberalism*, p. 14.
[25]Rawls, *Political Liberalism*, p. 66.

when these are well established and *not controversial*'.[26] If Rawls's principles depend on claims of 'science and social thought' that are contested by some reasonable people, then they cannot be expected to endorse these principles.

From the first-person perspective, the strategy of exclusion introduces a justificatory deficit. Given the exclusion of premises which may be relevant we may not be able rule out alternative views, and fail to identify a unique view as the correct one. We will be faced with what I called the weak views problem in the previous chapter. In 3.3, I illustrated this possibility with the disagreement between prioritarians and egalitarians, and with different measures of inequality. And more importantly, as we saw in 3.3, the strategy of exclusion can result in the false views problem. In 3.3 I illustrated this with the examples of Elif and Creon's edict. There will, therefore, be a gap between principles which may be legitimately imposed, and what it is rational, from the perspective of the individual inquirer, to believe about justice.

Rawls is not unaware of the problems with the strategy of exclusion. He addresses the problem in *Political Liberalism*. Rawls first formulates the problem we've looked at, though without, I believe, appreciating the extent of it:

[26]Rawls, *Political Liberalism*, p. 67 emphasis added. See also Rawls, *Political Liberalism*, p. 224-5 and Rawls, *Collected Papers*, p. 328-9. In contrast, in *Theory* Rawls introduced the information available to the denizens of the original position as follows: 'It is taken for granted, however, that they [the denizens of the original position] know the general facts about human society. They understand political affairs and the principles of economic theory; they know the basis of social organization and the laws of human psychology. Indeed, the parties are presumed to know whatever general facts affect the choice of the principles of justice.' Rawls, *Theory*, p. 505.

> 'How can it be either reasonable or rational, when basic matters are at stake, for citizens to appeal only to a public conception of justice and not to the whole truth as they see it? Surely, the most fundamental questions should be settled by appealing to the most important truths, yet these may far transcend public reason!'[27]

Rawls's response is that honoring public reason is required by the principle of liberal legitimacy. The value of legitimacy overrides other values, and honoring the limits imposed by public reason are a requirement of legitimacy. Rawls illustrates his point with an example:

> 'Why the apparent paradox of public reason is no paradox is clearer once we remember that there are familiar cases where we grant that we should not appeal to the whole truth as we see it, even when it might be readily available. Consider how in a criminal case the rules of evidence limit the testimony that can be introduced, all this to ensure the accused the basic right of a fair trial. Not only is hearsay evidence excluded but also evidence gained by improper searches and seizures, or by the abuse of defendants upon arrest and failing to inform them of their rights. Nor can defendants be forced to testify in their own defence. Finally, to mention a restriction with a quite different ground, spouses cannot be required to testify against one another, this to protect the great good of family life and to show public respect for the value of bonds of affection'[28]

[27] Rawls, *Political Liberalism*, p. 66.
[28] Rawls, *Political Liberalism*, p. 218.

It is true that improperly gained evidence cannot be introduced in a criminal case. However, our epistemic response to inadmissible evidence need not be the same as our practical response to inadmissible evidence. Suppose a juror has somehow received improperly gained evidence which gives her conclusive reason to believe that the defendant is guilty. In this case, she should believe that the defendant is guilty, even though she should not vote 'Guilty'.[29]

In reflecting on questions of justice, we are in the position of jurors who are in possession of inadmissible evidence. We have beliefs that are not shared by other reasonable people, and these beliefs have a role in determining what we ought to believe. As a result, there is a gap between what we are rationally required to believe, and, if the principle of liberal legitimacy is correct and trumps other considerations, what we are entitled to act on. This gap would not exist if the congruence requirement we looked at in the previous chapter applied to all of our disagreements with different reasonable people. The congruence principle required that we revise our opinions when others disagreed with us. However, we have seen that in most cases of disagreements about moral questions, we are rationally entitled to stick to our opinions in the face of disagreement.

Given the existence of this gap between the principles we are rationally required to believe, and the principles we are entitled to act on, we should interpret Rawls's claim that his constructivism in *Political Liberalism* relies on the method of reflective equilibrium with more care. A theory developed

[29]Actually, it is not at all obvious what the jury member should do in this case. We, surely, do feel the pull for a 'Guilty' vote in the case of serious crimes.

so that it can pass the test of liberal legitimacy is unlikely to be the best theory of justice from the individual inquirer's perspective. So it is unlikely to pass the test of reflective equilibrium as a theory of *justice*. Nevertheless, the theory which is designed with a view to the principle of liberal legitimacy can be in reflective equilibrium as a theory of all things considered judgements about the correct principles to put into practice, and, more specifically, about what would be legitimate. If this is to be so, we are nevertheless going to need to carry out justifications of the principles Rawls offers from the first-person perspective. We are going to need an account of why legitimacy trumps justice. If we go back to the example of the juror who has come by inadmissible evidence, the juror needs an account of why her verdict should not reflect inadmissible evidence, and in particular, why it is more important that rules of evidence are upheld even though this may result in justice not being done. And this account should be one which allows the individual to take into account all her relevant beliefs.

There are, then, three distinct tasks which we ought to distinguish between. Firstly, there is the task of coming up with a theory of justice that is best from the first-person perspective. Developing a theory that has this goal can make use of comprehensive philosophical doctrines and any other beliefs that are not shared by others. Secondly, there's the task of developing a set of principles that pass the test of liberal legitimacy. The derivation of these principles should follow Rawls's strategy of exclusion. As we have seen in the previous chapter, the fact that an argument doesn't rely on premises we deny and relies on premises we affirm doesn't entail that its conclusions

will be acceptable to us. Consequently, the results of the project aiming to come up with a theory of justice that is best from the first-person perspective and a theory that passes the test of liberal legitimacy can differ significantly. In such a case, there's a third task: coming up with an account, from the first-person perspective, of why legitimacy trumps justice and other values.[30]

4.5 Cohen on Justice and Other Virtues

In *Political Liberalism* Rawls supplements the strategy of exclusion with another argumentative strategy which he doesn't state explicitly. He relies on this strategy to overcome what I called the weak views problem in section 3.3. This strategy is to introduce premises which are widely shared but are not obviously relevant to our beliefs about justice. We can call this strategy the *strategy of inclusion*. For instance, no matter what disagreements people have about the correct metric of justice in theory, they can agree that relying on welfare as the metric of justice is extremely difficult, and perhaps impossible, to put into practice. Accordingly, even proponents of welfarist metrics of justice can settle down for primary goods as the metric we should rely on in practice. And this is one of the arguments Rawls offers in favour of primary goods. The list of primary goods are introduced 'to find a *practicable* public basis of interpersonal comparisons based on objective features of citizens' social circumstances open to view'.[31] Rawls is, however, aware that we can have beliefs about justice which are beyond what's practicable.

[30]Rawls calls this last process embedding the political conception of justice into one's comprehensive doctrine. See Rawls, *Political Liberalism*, p. 386

[31]Rawls, *Political Liberalism*, p. 181, emphasis added.

He grants that comprehensive moral doctrines do not need to follow this restriction. He writes: 'we must respect the constraints of simplicity and availability of information to which any practicable political conception (*as opposed to comprehensive moral doctrine*) is subject'.[32] For Rawls, then, it's not an objection to a comprehensive moral theory about justice that it's not practicable. The restriction applies to public conceptions of justice.

Bolstered with the strategy of inclusion, Rawls's strategy of exclusion may be able to identify a unique set of principles and overcome the weak views problem. However, these principles will not be acceptable as principles of *justice* to someone who has welfarist intuitions. The strategy of exclusion results in loss of data which from the personal perspective is relevant. And the strategy of inclusion results in the introduction of alien premises to the question of what justice is.

Cohen makes a similar criticism of Rawls's strategy of inclusion. Comparing my critique with his is will help me clarify my position. According to Cohen, Rawls's constructivism 'fails to distinguish between justice and other virtues'.[33] Cohen argues that this failure is due to the question that the agents of construction are set to answer. The agents of construction are 'not asked to say what justice is: it is we who ask ourselves that question, and the constructivist doctrine is that the answer to our question is the answer to the different question that is put to constructivism's specially designed selectors,

[32]Rawls, *Political Liberalism*, p. 181, emphasis added.
[33]Cohen, *Rescuing Justice*, p. 275.

which is, what are the optimal rules of social regulation?'[34] Since concerns other than justice enter into identifying the optimal rules of social regulation, the outcome of the constructivist procedure will not identify principles of justice, but will produce all-things-considered principles for the regulation of society. In the case of Rawls's constructivism, Cohen suggests, the deliberations of the denizens of original position will reflect a concern with stability of political arrangements, Pareto optimality, and publicity.[35] Cohen does not deny that these are legitimate concerns for institutional arrangements. His point is that these are concerns separate from justice.

Cohen's criticism is much stronger than my intended criticism. According to Cohen, the fact that considerations such as stability and publicity enter into the deliberations of the denizens of the original position entails that the principles which emerge out of the original position cannot be principles of justice. He takes this to be a conceptual point. For instance, when discussing stability, Cohen writes

> 'For to treat the evident desideratum of stability as a constraint on what justice might be thought to be, to judge that principles qualify as principles of justice only if, once instituted, their rule has a propensity to last, is absurd. It would mean that one could not say such entirely intelligible things as "This society is at the moment just, but it is likely to lose that feature very soon: justice is such a fragile achievement"; or "We don't want our society to be

[34]Cohen, *Rescuing Justice*, p. 275.
[35]Cohen, *Rescuing Justice*, p. 286.

> just only for the time being: we want its justice to last". It would mean that Plato was conceptually confused when he argued, on empirical grounds, in Book VIII of The Republic, that a just society was bound to lose its justice over time.'[36]

I believe Cohen intends this argument to be conclusive. However, if we take arguments of this type, which rely on what we are inclined to say, and take how we pre-theoretically understand different concepts we employ to be related as conclusive evidence, we would rule out the possibility of saying anything interesting about justice, or, for that matter, any other concepts. In order for our philosophical accounts to be interesting they need to be proposals which we are not pre-theoretically obvious. We might, nevertheless, take Cohen's point to establish a *prima facie* case against the inclusion of stability as a desideratum of justice. Accordingly, the relevance of stability to justice needs to be shown. On my account, the fact that a proposed principle would not be stable should not be allowed to defeat a principle that is proposed as a principle of justice. Nevertheless, if the original position produces intuitively compelling principles only if the concern with stability is built into the original position, this is not problematic. This would count as evidence that justice is related to stability despite initial appearances to the contrary.

It should be clear that my diagnosis of the source of the problem I've been discussing is different from Cohen's diagnosis. Cohen thinks that the concerns relevant to principles of regulation but alien to justice enter into

[36]Cohen, *Rescuing Justice*, p. 328.

Rawls's theory because of the original position. I have argued that these concerns enter into Rawls's theory due to Rawls's argumentative strategy of inclusion.

4.6 Publicity, Stability, and the Original Position

I have suggested that as long as the principles which emerge out of the original position are intuitively compelling the inclusion of concerns seen as irrelevant is not a problem. In this section, I shall argue that it is not inevitable that a concern with stability or publicity enters into the deliberations of the denizens of the original position. (I shall return to the role of Pareto in section 7.2.)

We should begin by observing that for Rawls stability does not play a role in the argument for the principles of justice from the original position. It is a further test Rawls imposes on the principles which emerge out of the original position. Rawls writes:

> '... [W]hat I have tried to show is that the contract doctrine is superior to its rivals on this score [stability], and therefore that the choice of principles in the original position need not be reconsidered.'[37]

Rawls would not need to show that the principles which emerge out of the original position are stable if a concern with stability already entered into the choice of principles in the original position.

[37] Rawls, *Theory*, p. 505.

It may be claimed that Rawls is mistaken about the reasoning in the original position, and given the way the original position is constructed, the denizens of the original position need to be concerned with stability. This challenge can be avoided by altering the design of the original position so that the denizens of the original position cannot take stability into consideration. One possible change is to stipulate that the parties assume that all the principles they choose are equally stable. If they are deliberating under this assumption, then, considerations of stability cannot enter into their deliberations. Instead of stipulating that all of the principles the denizens of the original position are considering are equally stable, we can thicken the veil of ignorance. This way the denizens of the original position would lack the information necessary for evaluating the stability of the principles they are considering. Both of these alterations can ensure that the deliberations of the denizens of the original position are not influenced by stability concerns. Accordingly, it's not inevitable that a concern with stability will enter into the deliberations of the denizens of the original position.

In order to assess Cohen's objection that Rawls's constructivist procedure inevitably introduces a concern with publicity, we need to distinguish between two different senses of publicity. In *Theory*, Rawls introduces publicity as formal constraint on the concept of right. The idea is that the principles of justice which are evaluated by the denizens of the original position are assumed to be public knowledge and followed by everyone.[38] Rawls uses this

[38]Rawls, *Theory*, p. 115. Note that what has to be public knowledge is the fact that these principles are being followed and not the level to which they have been realized. This is clear from Rawls's concession that 'The application of the difference principle in a precise way requires more information than we can expect to have'. See Rawls, *Theory*,

condition to rule out conceptions of justice which work only 'if understood and followed by a few or even all, so long as this fact were not widely known' such as government house utilitarianism.[39] Andrew Williams's publicity condition, which is the subject of much of Cohen's discussion of publicity in *Rescuing Justice and Equality*, is not the same as the publicity constraint Rawls uses in *Theory*. Williams develops the publicity requirement from Rawls's discussion of the basic structure as a public system of rules, and Rawls's definition of a public system of rules. Williams holds that the principles of justice are public when what the principles of justice are, what the principles of justice demand in particular cases, and the extent to which the demands of justice are satisfied are known by all concerned.[40]

It is not clear whether the denizens of the original position should be concerned with publicity as either Rawls or Williams conceive of it. The publicity condition as Rawls defines it limits the principles proposed for the consideration of the denizens of the original position. It is not a property of principles they seek, but one which narrows down the principles that they consider in the original position. Furthermore, publicity, as Rawls understands it, is a reasonable demand relevant to justice. It is plausible to claim that citizens do have a *right* to know about the principles which govern their society. So, a concern with publicity as Rawls understands it would not be a consideration obviously alien to justice.

p. 174.

[39] Rawls, *Theory*, p. 398.

[40] Andrew Williams, 'Incentives, Inequality, and Publicity', *Philosophy and Public Affairs*, 27 (1998):3, p. 233.

Even though the denizens of the original position are equipped with certain facts, they lack the information that they would need for deciding how the principles they have chosen are to be implemented. This would rule out the denizens of the original position being concerned with Williams's publicity condition. To ascertain how the principles chosen in the original position are to be implemented Rawls envisages a four-stage sequence where the limitations on the knowledge available to the denizens of the original position are lifted in stages. When they are choosing the principles of justice, 'the only particular facts known to the parties are those that can be inferred from the circumstances of justice'.[41] They also know the 'first principles of social theory', by which I understand facts that are true for all societies, but they do not know which kinds of societies presently exist. As a result, they lack the information that would be required to assess whether the principles they have chosen would be able to satisfy Williams's publicity condition. For instance, informational constraints we suffer from today may make it impossible for certain principles to satisfy Williams's publicity condition. However, these constraints may disappear in the future thereby making it possible for these principles to satisfy Williams's publicity condition. Since the denizens of the original position do not know specific facts about their society or its place in history, they cannot deliberate about whether the principles they choose satisfy Williams's conditions. This is true unless it is impossible for certain principles to satisfy the publicity condition in all human societies under the circumstances of justice.

[41]Rawls, *Theory*, p. 175.

The possibility that some principles of justice may be unable to satisfy Williams's publicity condition in all human societies leaves room for the denizens of the original position to deliberate about publicity. But do they have reason to be concerned with it from their perspective? Williams's publicity requirement can be seen as a necessary condition for people in a society to be assured that the others are following the principles of justice. We can be assured of others' compliance only if what justice demands in specific cases and whether everyone is doing their bit can be known publicly. The assurance condition plays a crucial role in contract doctrines, because actual agreement and compliance of the relevant parties is a precondition for the principles of justice. However, Rawls's doctrine is not an actual contract doctrine. It's a hypothetical contract doctrine.[42] Furthermore, in evaluating the principles that are presented to them, the denizens of the original position assume strict compliance; 'the principles of justice are chosen on the supposition that they will be generally complied with'.[43] Accordingly, the problem of assurance does not enter into their deliberations. They are deliberating about which principles are rationally advantageous for them when strict compliance holds. Therefore, worries about assurance and the related question of publicity as Williams presents it do not enter into their deliberations.

To conclude, the requirement of Rawlsian publicity comes prior to the presentation of principles to the denizens of the original position, and it is

[42]This statement needs to be qualified. Rawls has several elements in his arguments that take it beyond a hypothetical contract. See, in particular, Section 29 of Rawls, *Theory*. However, the original position, which is the focus discussion, involves a hypothetical contract–if it involves a contract at all.

[43]Rawls, *Theory*, p. 215.

a concern that is *prima facie* relevant to justice. With regard to publicity as Williams conceives of it, the denizens of the original position do not take it into consideration because (a) they lack much of the information which would be necessary to evaluate whether the principles they adopt would satisfy Williams's publicity condition, and (b) since the original position works under the assumption of strict compliance and is not an actual contract doctrine, they do not have a reason to be concerned with assurance.

Finally, there is another publicity related notion in Rawls's theory which plays a central role in *Political Liberalism*, and ties in with my argument that the shortcomings Cohen identifies stem from Rawls's argumentative strategy. This notion is the idea that a political conception of justice has a public role in a well-ordered society.[44] According to Rawls:

> '... [T]he public role of a mutually recognized political conception of justice is to specify a point of view from which all citizens can examine before one another whether or not their political institutions are just. It enables them to do this by citing what are recognized among them as valid and sufficient reasons singled out by that conception itself.'[45]

The public role of a political conception of justice requires that (a) it is public knowledge that the 'public principles of justice' are accepted and recognized; (b) citizens know that the basic structure is just 'on the basis of commonly

[44]The public role of justice, and what Rawls believes follow from it, are similar to Williams's publicity requirement.

[45]Rawls, *Collected Papers*, p. 426.

shared beliefs confirmed by methods of inquiry and ways of reasoning generally accepted as appropriate for questions of political justice'; and (c) the first principles of justice are accepted in light of 'general beliefs about human nature and the way political and social institutions generally work' that reflect 'the current public views'.[46] Here we find grounds for Rawls's argumentative strategies. The need for public justification of the theory required the formulation of the theory in a way that could be the subject of an overlapping consensus. This, we saw, required the strategy of exclusion. Another requirement which stems from the public role of a political conception of justice is its ability to be openly checkable. Citizens must be able to *show* to one another that their political institutions are just. This requires 'the constraints of simplicity and availability of information to which any practicable political conception. . . is subject'.[47] This is part of what I identified as Rawls's strategy of inclusion. Publicity, so understood, is a significant virtue of social institutions. However, it should be clear from our discussion of Rawls's argument in *Political Liberalism* in section 4.4 that it is a consequence of Rawls's commitment to the liberal principle of legitimacy and the argumentative strategy he employs rather than a consequence of constructivism and the original position.

4.7 Conclusion

In this chapter, I've identified the version of constructivism that I will defend in the next chapter: bound and conservative constructivism about justice as

[46]Rawls, *Collected Papers*, p. 66-7.
[47]Rawls, *Political Liberalism*, p. 182.

it is presented in *Theory*. The fact that Rawls's constructivism rests on a normative basis and seeks to conserve our existing beliefs about justice, will be crucial to my case in the next chapter that Rawls doesn't need to deny Cohen's claim about the relationship between facts and principles.

I argued that Rawls's constructivism in *Theory* and his constructivism in *Political Liberalism* have different goals. The first seeks to offer an account of justice that is best from the first-person perspective. The second project seeks to offer a set of principles that pass the test of liberal legitimacy. The second project needs to employ an argumentative strategy that requires discarding data which from the first-person perspective is relevant to thinking about justice, and introducing data which from the first-person perspective is *prima facie* irrelevant to thinking about justice.[48] For this reason, the constructivism in *Political Liberalism* is likely to fail as an account of justice that is best from the first-person perspective. However, this argumentative strategy is not definitive of constructivism, and Rawls's constructivism in *Theory* does not follow this strategy.

I argued that it is the argumentative strategies that Rawls employs in *Political Liberalism*, and not the original position, or constructivism, that gives rise to Cohen's criticism that Rawls's theory can't distinguish between justice and other virtues. Freed from Rawls's argumentative strategy one can alter the design of the original position in ways that allows one to remove

[48]The qualification *prima facie* is required, because, as I argued on page 107, once the whole theory of justice is worked out we may find that the data, which pre-theoretically seemed irrelevant, was relevant.

considerations alien to justice from the deliberations of the original position.

My focus on the constructivism of *Theory* has one important consequence for my discussion of egalitarian ethi in chapters 7 and 8. I believe that, in the absence of a concern with publicity and practicality, welfare remains a plausible candidate for the metric of justice. Consequently, when I examine the case for egalitarian ethi, I shall examine the case for them both on the assumption that welfare is the metric of justice, and on the assumption that primary goods are the metric of justice.

Finally, I should emphasize that my criticism of Rawls's argumentative strategy in *Political Liberalism* is not intended to establish that the account Rawls offers there fails on its own terms, that is, as an account of principles that may be legitimately imposed. All that my argument shows is that we have good reason to expect that the principles of justice and the principles that may be legitimately imposed diverge.

Chapter 5

Rawls and Cohen on Facts and Principles

5.1 Introduction

In the previous chapter, I identified Rawls's constructivism as an example of bound and conservative constructivism. In this chapter, I defend Rawls's constructivism against Cohen's critique of it in chapter 6 of his *Rescuing Justice & Equality*, which is a reproduction of his his article 'Facts and Principles'.

Cohen has argued that

> (C): '. . . [A normative] *principle can reflect or respond to a fact only because it is also a response to a principle that is not a response to a fact.* To put the same point differently, principles that reflect facts must, in order to reflect facts, reflect principles

that don't reflect facts.'[1]

The terms used in C are defined as follows. A normative principle (following Cohen's usage hereafter referred to, for short, as a 'principle') is 'a general directive that tells an agent what (they ought, or ought not) to do'.[2] A fact is 'any truth *other than (if any principles are truths) a principle*'.[3]

Cohen gives the following example to illustrate C. Suppose someone affirms the principle '*we should keep our promises* (call that P) because *only when promises are kept can promisees successfully pursue their projects* (call that F)'.[4] Cohen argues that this person has to concede that she affirms the further principle P_1 which says that 'we should help people *to pursue their projects*'.[5] It is P_1 that makes F a ground for P. Fact-sensitive principles are principles which have facts among the grounds for affirming them.[6] It is Cohen's thesis that the scrutiny of fact-sensitive principles will always reveal a deeper fact-insensitive principle that explains why a fact grounds a certain principle. According to Cohen, Rawls denies C. Furthermore, the Rawlsian construction is inevitably fact dependent. As Rawls concedes, the denizens of the original position (hereafter the OP) cannot make a decision in the absence of facts.[7] Since according to C, someone who affirms a principle in light of facts is committed to another principle, Rawlsians are mistaken about

[1]G. A Cohen, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008), p. 232.
[2]Cohen, *Rescuing Justice*, p. 229, emphasis in the original.
[3]Cohen, *Rescuing Justice*, p. 229, emphasis in the original.
[4]Cohen, *Rescuing Justice*, p. 234, emphasis in the original.
[5]Cohen, *Rescuing Justice*, p. 234, emphasis in the original.
[6]Cohen, *Rescuing Justice*, p. 231.
[7]John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 138.

the structure of their beliefs. There is a fact-insensitive principle Rawlsians are committed to but have failed to explicate.

I shall argue that Rawls does not deny C, but argues for an altogether different thesis. To distinguish Cohen's thesis from Rawls's thesis, I'd like to introduce some terminology. A principle is *metaphysically fact-insensitive* if its truth doesn't depend on any facts. A principle is *epistemically fact-insensitive* if our belief in it doesn't depend on knowledge of facts. Principles, such as Rossian prima facie duties, which we are inclined to affirm without knowledge of facts are examples of epistemically fact-insensitive principles. Cohen's central claim is that there are metaphysically fact-insensitive principles. I shall argue that Rawls's central claim is that metaphysically fact-insensitive principles and epistemically fact-insensitive principles need not coincide, and, in fact, epistemically fact-insensitive principles can be metaphysically fact-sensitive. I shall also argue that Rawls acknowledges the existence of principles which underlie the OP, but these principles are not principles of distributive justice.

Rawls's constructivism and his comments on the basis of which Cohen argues that Rawls denies C are best understood and evaluated within a general framework of moral theorizing. This is the framework we developed in Chapter 2. In Section 5.2, I offer an alternative interpretation of the passages from Rawls that Cohen uses in arguing that Rawls denies C. Section 5.3 develops this interpretation further and defends Rawls's claims about the relationship between facts and principles. I argue that what Rawls rejects

are approaches that rule out the use of facts in the derivation of higher-level moral principles from lower-level ones, and the use of facts in justifying a theory to us. Rawls, like many utilitarians, denies that it is an objection to a moral theory that it fails to capture our intuitions about cases which cannot arise in our world. This is not because a theory's failure to capture such intuitions doesn't make a practical difference, but because Rawls, like some utilitarians, refuses to take such intuitions at face value. It is possible that even though Rawls does not deny C, the OP only has force if C is false. Therefore, a separate defence of the OP and the use of facts in it are necessary. This task is taken up in Section 5.4. I argue that Rawls does not deny that metaphysically fact-insensitive principles lie behind the OP and that the use of facts in the OP does not entail that the principles which emerge out of the OP are epistemically fact-sensitive principles.

Three caveats. My defence of Rawls's constructivism applies only to the version of constructivism in *Theory*. As we've seen in Chapter 4, in later works, the role of constructivism within Rawls's theory, and the questions Rawls addresses in his theory change substantially. The version of constructivism I defend is also the one Cohen takes issues with. Secondly, my defence is limited to what I in the previous chapter called bound and conservative constructivism. Thirdly, my goal is not to defend the content of Rawls's theory, but to defend its structure, which is what is at stake in Cohen's critique. Accordingly, I assume that Rawls's derivation of his principles from the OP is correct, and assume that Rawls's theory has the theoretical virtues which I argued in 2.3 we require from moral theories.

5.2 Does Rawls Deny Cohen's Thesis?

Rawls's most significant remarks concerning the employment of facts in moral theories, which are also cited by Cohen to argue that Rawls denies C, occur in his response to a possible objection to the difference principle. The objection is that the difference principle makes 'the justice of large increases or decreases in the expectations of the more advantaged depend upon small changes in the prospects of the worst-off'.[8] For instance, the difference principle may allow huge inequalities if they are necessary for improving the condition of the worst-off even though the improvements in how the worst-off fare are infinitesimally small. Rawls responds to this objection by arguing that such cases will not arise. In defence of his use of a factual claim he says:

> 'We should also observe that the difference principle not only assumes the operation of other principles, but it presupposes as well a certain theory of social institutions. In particular... it relies on the idea that in a competitive economy (with or without private ownership) with an open class system excessive inequalities will not be the rule... Now the point to stress here is that there is no objection to resting the choice of first principles upon the general facts of economics and psychology. As we have seen, the parties in the original position are assumed to know the general facts about human society. Since this knowledge enters into the premises of their deliberations, their choice of principles is relative to these facts. What is essential, of course, is that these premises be true

[8] Rawls, *Theory*, p. 136.

> and sufficiently general. It is often objected, for example, that utilitarianism may allow for slavery and serfdom, and for other infractions of liberty. Whether these institutions are justified is made to depend upon whether actuarial calculations show that they yield a higher balance of happiness. To this the utilitarian replies that the nature of society is such that these calculations are normally against such denials of liberty.
>
> Contract theory agrees, then, with utilitarianism in holding that the fundamental principles of justice quite properly depend upon the natural facts about men in society. This dependence is made explicit by the description of the original position: the decision of the parties is taken in the light of general knowledge.'[9]

In this passage, Rawls conflates two different perspectives. He shifts back and forth between our perspective, that is, the perspective of someone trying to decide whether the principles which emerge out of the OP formulate her convictions, and the perspective of the denizens of the OP. In mentioning the fact that there will not be excessive inequalities in competitive economies, Rawls is showing that certain distributions which we would consider unjust will not be considered just by the difference principle. He then mentions that the choice of the denizens of the OP is sensitive to facts. However, the denizens of the OP do not assume that such inequalities will not be the rule in order to choose the difference principle. In fact, even if they do not make Rawls's empirical assumption, they would still adopt the difference principle.

[9]Rawls, *Theory*, p. 137.

If maximin is the correct decision rule in the OP and the denizens of the OP are not motivated by envy, then, they have no reason to object to inequalities which may arise even if they do not assume that there will not be excessive inequalities in competitive economies. It is us evaluating the principles the denizens of the OP have chosen who have to make this empirical assumption so that we are assured the difference principle does not give rise to inequalities we find objectionable.[10]

The agreement between contract theory and utilitarianism about the role of facts, which Rawls refers to on the passage I've cited, cannot be about the use of facts in the OP because utilitarianism, with the exception of Harsanyi's version of it, does not use such a device. Furthermore, utilitarians don't think that their principle is made true by certain facts obtaining. The parallel Rawls draws in the passage quoted isn't between his two principles of justice and utilitarianism, but between contract theory and utilitarianism. According to both contract theory and utilitarianism, principles of distributive justice depend upon facts. Standardly, utilitarians don't offer utilitarianism as a principle of justice. Nevertheless, they have an account of justice. Rawls outlines the utilitarian strategy of accounting for principles of justice, which is based on Mill's account, as follows:

> '... [F]rom a utilitarian standpoint the explanation of these precepts [of justice] and their seemingly stringent character is

[10] A concern with the social bases of self-respect, which Rawls considers a primary good, might give rise to concern with inequalities in income. However, Rawls's discussion in this part of the text doesn't refer to it. Accordingly, it can be safely set aside when interpreting his discussion.

> that they are those precepts which experience shows should be strictly respected and departed from only under exceptional circumstances if the sum of utility is to be maximized. Yet, as with all other precepts, those of justice are derivative from the one end of attaining the greatest balance of satisfaction.'[11]

Rawls's parallel strategy is to argue that principles of distributive justice are the principles which would be chosen in the OP in light of facts, and that our common sense precepts of justice and beliefs about the stringency of demands of justice are accounted for by contract theory.[12]

Contract theory and utilitarianism agree that we can assume certain facts hold when testing the principles they have proposed against our intuitions or in deriving higher-level principles from lower-level ones. They both agree with a theoretical strategy which says: 'Here are the fundamental principles of justice: When conjoined with facts about the world, they will generate your evaluative judgements, or evaluative judgements you will find acceptable upon due reflection. It is irrelevant whether they will fail under factual assumptions which do not actually hold.'[13] As I shall elaborate below, both utilitarianism and Rawls claim that moral beliefs, such as our beliefs about distributive justice, many of which are epistemically fact-insensitive, can be metaphysically fact-sensitive.

[11] Rawls, *Theory*, p. 23.

[12] Rawls, *Theory*, pp. 25 and Chapter 5.

[13] This strategy, which many utilitarians have employed, has been articulated by R.M. Hare. See R. M. Hare, 'Ethical Theory and Utilitarianism', in: Amartya Sen and Bernard Williams, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982), p. 30. For an application of this strategy to the question of slavery, see R. M. Hare, 'What is Wrong with Slavery', *Philosophy and Public Affairs*, 8 (1979):2.

For the sake of completeness, I'd like to offer an interpretation of the other passage from Rawls which Cohen uses to bolster his claim that Rawls denies C, even though this is not pertinent to the role of facts in Rawls's theory which will be the subject of our discussion in the remaining sections. The sentence Cohen quotes reads: 'Conceptions of justice must be justified by the conditions of our life as we know it or not at all'.[14] Read on its own, it seems to support Cohen's understanding of Rawls. However, when the sentence is read in context, it lends itself to a different reading. Rawls writes:

> 'Conceptions that might work out well enough if understood and followed by a few or even by all, so long as this fact were not widely known, are excluded by the publicity condition. We should also note that since principles are consented to in the light of true general beliefs about men and their places in society, the conception of justice adopted is acceptable on the basis of these facts. There is no necessity to invoke theological or metaphysical doctrines to support its principles, nor to imagine another world that compensates for and corrects the inequalities which the two principles permit in this one. Conceptions of justice must be justified by the conditions of our life as we know it or not at all.'[15]

At the end of the last sentence, there is a footnote, which states that this rules out things like Plato's noble lie and the advocacy of a religion that is

[14]Rawls, *Theory*, p. 398. Cohen quotes this sentence on Cohen, *Rescuing Justice*, pp. 231-2, n. 4.

[15]Rawls, *Theory*, p. 398.

believed to be false to buttress a social system. This is not a denial of C. Rawls is saying that if false assumptions are necessary for principles to be justified to those governed by them, then, this is unacceptable. And if false assumptions are needed for the denizens of the OP to choose a principle, then, this is also unacceptable. What Rawls says there may commit him to the claim that principles may depend on facts, and that principles of distributive justice do, in fact, depend on facts, but Rawls clearly is not arguing that all principles are fact-sensitive. What Rawls says here seems to be nothing more than 'No sound principle is grounded in false beliefs' rather than 'All principles are grounded in facts'.

5.3 The Role of Facts in Theorizing

Cohen is interested in the logical relations between facts and principles. Accordingly, he assumes that the interlocutors in the examples he uses to illustrate his thesis have a clear grasp of what their principles and the grounds for these principles are.[16] Cohen's interlocutors have, as it were, carried out their theorizing and have a view regarding the structure of moral reality. Cohen is defending the thesis that there are metaphysically fact-insensitive principles. Rawls, however, is defending a thesis about our theorizing. Rawls's thesis is that when theorizing we can assume that certain facts hold, and principles which are epistemically fact-insensitive need not be metaphysically fact-insensitive.

[16]Cohen, *Rescuing Justice*, p. 233.

If we return to Cohen's example I cited in 5.1, we are to imagine someone who affirms the principle we should keep our promises (P) because '*only when promises are kept can promisees successfully pursue their projects*' (F). Given her previous reason, this person, as Cohen argues, has to also affirm the principle 'we should help people *to pursue their projects*' (P_1). F's obtaining *makes* P true. Someone who makes this claim commits herself to affirming P_1, which is a higher-level principle. To illustrate Rawls's position, I would like to examine the point of view of someone in the process of moral theorizing, and examine the possible ways in which P, F and P_1 may be related in the belief structure of a typical inquirer who is in the process of theorizing.

In 2.3, I suggested that our aim in theorizing is to develop a unified account of the moral beliefs we hold at different levels of generality. On this account, the support for holding P may be the generalization from the observation of the wrongness of individual instances of breaking promises to P, the independent plausibility of P, or its derivation from higher-level principles. In the case of a typical inquirer, the main support for P comes from its apparent plausibility. If this is the case, what is the function of P_1? P_1 is a higher-level principle than P. Therefore P_1 accounts for more evaluative judgements. P_1 may play a role in interpersonal justification or a justificatory role in the beliefs of a single individual. To see the role of P_1 in interpersonal justification imagine we are trying to convince someone that we should keep our promises. This person holds the set of moral beliefs E. E does not contain any judgements regarding promising. Thus it does not contain P. Suppose

further that, the evaluative judgements in E are best accounted for by P_1. By showing that P is entailed by P_1, which accounts for this person's other judgements, we draw support for P. This may lead a person who holds E to revise his principles and come to hold P and P_1.

P_1 may also play a justificatory role for a single individual. A person who already holds P may come to hold P_1 given F and P_1's ability to account for other lower-level principles this person holds. A person who holds both P and P_1 independently will be more confident of his principles if F holds. Similarly, a person who holds only P_1 may come to hold P when F obtains. *The transfer of conviction* in this account is multi-directional. We may come to hold lower-level principles because they are entailed by higher-level principles we hold, we may come to hold higher-level principles because they account for our lower-level principles, or we may become more confident of our principles because they mutually support each other.

Given that the transfer of conviction is multi-directional, it is difficult to identify which principles are epistemically fact-sensitive and which ones are not. Suppose a person finds lower-level principles P_2, P_3 and P_4 plausible. There is a higher-level principle, P_5, that she also finds plausible. P_5 accounts for P_2, P_3 and P_4 when F holds. If she believes that F holds, then her confidence in P_2, P_3, P_4 and P_5 will have increased. In one sense all of these principles are epistemically fact-insensitive, because she holds them without reference to facts. However, her increased confidence in them is due to facts. It is also possible that P_5 conflicts with one of the lower-level principles when

not-F holds. In such a case, the set containing P_2, P_3, P_4 and P_5, or to be more precise the consistency of this set of beliefs, would be epistemically fact-sensitive even though the principles by themselves are not.

Alternatively, she may find P_5 neither plausible nor implausible on its own, but come to hold it given its ability to account for P_2, P_3 and P_4. Sometimes we arrive at our higher-level principles inductively. We move up to higher-level principles when we are able to subsume our lower-level principles under these higher-level ones. This is possible only when we establish conceptual connections or facts enable us to subsume our lower-level principles under higher-level principles. Insofar as our higher-level principles are not given, we may need facts to develop them from lower-level ones. If higher-level principles are arrived at inductively by relying on facts, they will be epistemically fact-sensitive. (An example might clarify this claim. Suppose that the truth of utilitarianism is self-evident. In that case, we will not need to rely on facts. Now suppose that utilitarianism is not self-evident, but it is nevertheless true. How could we come to know it? One way would be to argue that it is the best explanation of our lower level principles, and other intuitive judgements that we make. In order to make this argument, we would need to rely on facts. The argument would roughly be that given facts about the world, these principles and intuitive judgements are best explained by utilitarianism. This argument would have to rely on facts, because it would need to establish that our lower level principles are conducive to the maximization of utility.) This is another way of construing the Rawlsian reliance on facts and the reliance on facts that, according to Rawls, exists in utilitarianism.

Rawls and utilitarianism both agree that (a) we can use facts in moving from lower-level principles to higher-level ones, (b) we can rely on facts in ensuring a match between higher and lower-level principles; and (c) we can hold epistemically fact-sensitive sets of principles.

It is worth reiterating that I am not denying the existence of metaphysically fact-insensitive principles. The point is rather that there is, often, a mismatch: Metaphysically fact-insensitive principles will tend to be higher-level and epistemically fact-sensitive, whereas epistemically fact-insensitive principles will usually be lower-level principles, and not metaphysically fact-insensitive. It is this thesis which Rawls and utilitarians, as Rawls interprets them, hold.

Suppose the account of promising in Cohen's example is correct.[17] Then the transfer of conviction in the case of a typical inquirer would not correspond to the structure of moral truth. Speaking metaphorically we can say that truth flows from higher-level principles to lower-level principles whereas conviction flows in both directions, and often from lower-level principles to higher-level ones. Usually, our theorizing begins from lower-level principles and intuitive judgements. We arrive at higher-level principles by employing facts. Furthermore we are often more confident of lower-level principles than we are of higher-level principles. Consequently, if it was shown that promises do not necessarily contribute to people's successful pursuit of their projects, we would remain committed to P, and withhold judgement on P_1.

[17] Cohen isn't committed to claiming that this is the correct of account of promising. Cohen, *Rescuing Justice*, p. 234.

The role of facts in theorizing I've attributed to Rawls is exemplified by Brad Hooker's version of rule-consequentialism. Hooker argues for his version of rule-consequentialism by claiming that it can account for our lower-level principles.[18] This argument relies on facts, and we will hold the principles he proposes in light of facts. However, he claims that it is his theory of rule-consequentialism and not our lower-level principles which hold in all possible worlds.[19] In other words, Hooker claims that it is the epistemically fact-sensitive higher-level principles he proposes which are metaphysically fact-insensitive. Rawls wants to remain as non-committal as possible on metaphysical questions. Nevertheless he gestures towards a strategy similar to Hooker's, but then takes it back. He suggests that even though the principles which emerge out of the OP are contingent since they depend on facts, the conditions the OP embodies *may* be held to be necessary truths, but -here is the part he takes it back- it is better to regard them 'simply as reasonable stipulations'.[20]

Should we prefer a theory that is less reliant on facts in capturing our lower-level principles and intuitions about specific cases? I believe there is no clear cut answer to this question. There is a variety of considerations that have to be taken into account. For instance, we need to consider how certain we are of the facts that have been employed. We aren't likely to have qualms

[18]Brad Hooker, *Ideal Code, Real World: A Rule-Consequentialist Theory Of Morality*, (Oxford: Clarendon Press, 2000).

[19]Brad Hooker, 'Reflective Equilibrium and Rule Consequentialism', in: Brad Hooker, Elinor Mason and Dale E Miller, editors, *Morality, Rules, and Consequences: A Critical Reader*, (Edinburgh: Edinburgh University Press, 2000), p. 233.

[20]Rawls, *Theory*, p. 506. Note that if, as Cohen argues, Rawls denied C the natural thing for him to say here would be that there aren't any moral principles that are necessary truths.

about employing facts we are highly certain of in our theorizing. Another relevant point of comparison between a theory which does not employ facts and one which does is how well our goals in theorizing are served by these theories. We need to ask whether the theory in question provides responses to difficult cases, or whether it gives us a more systematic account of our moral beliefs. We also need to consider the scope of the facts we employ in our theorizing. Some facts about human nature and society apply only during certain periods of human history. We might hesitate to employ such facts in our theorizing. However, reliance on facts which have characterized human life and society in all epochs does not seem problematic. We can perhaps impose a more general requirement. Using the language of possible worlds, we can say that we do not require our moral theories to capture our intuitions about all possible worlds, but to capture our intuitions about a range of possible worlds that are close to ours.[21] In short, there is no answer prior to specific theories that suggests we should prefer less fact-dependent theories.

Imagine this scenario. One philosopher says to her opponent, 'Yes, under the usual conditions of human life, your theory captures and systematizes *all* of our moral beliefs, and provides illuminating answers in difficult cases. My theory accounts only for our intuitions about promises. But your theory can't capture our intuitions about cases regarding promises made in a world in which people's memory only reaches back to the previous week.' That this theory has a claim to capture our intuitions about cases which may arise in

[21] On the notion of the closeness of possible worlds see David K. Lewis, *On the Plurality of Worlds*, (Oxford: Blackwell, 2001).

all possible worlds, or at least in more possible worlds than its rival, does not seem to be a reason which commends it. Furthermore, given the success of the more general but fact-dependent theory, we are likely to doubt the reliability of our intuitions about promises in a world so different from ours.

The method of deriving higher-level principles from more specific principles or evaluative judgements through the employment of facts is consistent with the existence of higher-level principles we already have. As we have seen, Rawls rightly maintains that we have considered judgements at all levels of generality among which we seek reflective equilibrium.[22] The picture is one where principles at various levels of generality mutually support each other, and facts are brought into the mix. A procedure whereby we derive *all* of our specific judgements from a few higher-level principles without employing any facts would be needed to show the superiority of theorizing which does not rely on facts. Alternatively, these higher-level principles would have to be strong enough to override our lower-level principles and intuitive judgements. There is no way of ruling out the possibility of discovering such principles. However, there are no likely candidates at present. Alternatively the project of theorizing as I presented it in Chapter 2 may be rejected. This would be an impoverishment of moral and political philosophy which few would find palatable.

This account of the relationship between facts and principles sheds light on another passage quoted disapprovingly by Cohen. Rawls writes:

[22]John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 289.

> Some philosophers have thought that ethical first principles should be independent of all contingent assumptions, that they should take for granted no truths except those of logic and others that follow from these by an analysis of concepts. Moral conceptions should hold for all possible worlds. Now this view makes moral philosophy the study of the ethics of creation: an examination of the reflections an omnipotent deity might entertain in determining which is the best of all possible worlds. [23]

In this passage Rawls probably has Leibniz's ethics in mind.[24] The idea in Leibniz is that since God is omniscient and omnipotent he would create the best of all possible worlds. The ethics of creation is the inquiry into the principles which would have guided God's creation and his choice of the best of all possible worlds. The study of ethics as Rawls conceives it is an inquiry into the principles that underlie our evaluative judgements that we make in this world. Rawls is not denying that there are metaphysically fact-insensitive principles. He only claims that we can take facts as part of the data for our theories. What Rawls's theory seeks to explain is our evaluative judgements we make in this world, and accordingly facts can be used in our theorizing. His theory does not seek to capture our moral judgements about cases which could not arise in our world. If our higher-level principles were given, or derived solely by establishing conceptual links between them and our lower-level principles and evaluative judgements, and if these judgements and lower-level principles were either self-evident, or somehow reflected some

[23]Rawls, *Theory*, p. 289. Quoted in Cohen, *Rescuing Justice*, p. 261.

[24]John Rawls, *Lectures on the History of Moral Philosophy*, (Cambridge, Mass: Harvard University Press, 2000), p. 107-9.

timeless and unchanging truth, then they would hold in all possible worlds. This is what is meant by the reference to possible worlds.[25]

5.4 Cohen's Thesis and the Original Position

In the previous two sections, I argued that Rawls does not deny C and defended the way he conceives of the relationship between facts and principles. Even if this defence is successful the truth of C might threaten the conceptual coherence of the OP. To defend the structure of Rawls's theory we need to show that no such threat exists. In this section, I shall argue that constructivist procedures which are reliant on facts can capture epistemically fact-insensitive principles and the principles which underlie the OP are distinct from the principles which emerge from it.

It is true that the denizens of the OP cannot make any choices in the absence of facts. However, this does not mean that a fact-dependent constructivist procedure like the OP cannot capture our epistemically fact-insensitive principles. This point can be illustrated with Kant's categorical imperative procedure. The categorical imperative is a constructivist procedure. It asks

[25]It is worth noting that in the quoted passage Rawls's comments concern how we conduct our enquiries and not the nature of moral truth. We should also note the historical context of these passages. Firstly, Rawls is arguing against the then dominant approach in moral philosophy which primarily relied on linguistic analysis. Secondly, *Theory* was written before Kripke's arguments in *Naming and Necessity*, which distinguished between necessity, a priority, and analyticity, were widely known. This, I think, accounts for Rawls quickly moving from claims about conceptual analysis to talk of possible worlds in the passage quoted. For the argument that necessity, a priority, and analyticity were typically taken to be the same notion for much of 20^{th} century analytic philosophy prior to Kripke's work see Scott Soames, *Philosophical Analysis in the Twentieth Century*, Volume 1, (Princeton, N.J: Princeton University Press, 2003).

whether a certain maxim can be universalized. In thinking whether a maxim can be universalized we employ our knowledge of facts. For instance, when the procedure is applied to promising, it rejects a maxim such as 'I am to make a deceitful promise whenever it serves my purpose'. The test of this maxim assumes that people learn from past experience, and remember the past.[26] In this way, the procedure is fact-sensitive. We may, nonetheless, hold its outcome without reference to any facts as an epistemically fact-insensitive principle. I believe for most people the duty not to make false promises is one that they hold independently of facts and is one of the starting points in moral theorizing. It is for most of us an epistemically fact-insensitive principle that is captured by a fact-sensitive constructivist procedure.[27]

It will be asked why we should bother with the constructivist procedure if it all depends on whether its outcome matches one's considered convictions. In response: the constructivist procedure is not required to capture all of our intuitions. In fact, it is expected that in some cases when the principles which emerge out of the constructivist procedure conflict with principles we currently hold, it is the principles that emerge from the constructivist procedure that we should come to hold. Nevertheless, the constructivist procedure has to be able to capture our strongly held convictions which we

[26]Rawls, *Lectures on the History of Moral Philosophy*, p. 171.

[27]On some interpretations, the Categorical Imperative procedure allows us also to maintain that there are things which are wrong in all possible worlds. For instance, slavery would be wrong in all possible worlds. O'Neill argues that we can't will the maxim of becoming a slave as a universal law. She writes: '... [I]f everybody became a slave, there would be nobody with property rights, hence no slaveholders, and hence nobody could become a slave.' Similarly, we can't will the maxim of becoming a slave holder. Onora O'Neill, *Constructions of Reason: Explorations of Kant's Practical Philosophy*, (Cambridge: Cambridge University Press, 1989), p. 96.

may hold independently of facts. For this reason, it was necessary to show that the dependence of constructivist procedures on facts does not entail that the principles they produce will be epistemically fact-sensitive principles.

The OP itself also has to be acceptable to us and play an explanatory role if it is going to play a significant role in our theorizing, and make us give up some of the principles we hold prior to theorizing. At this point a second objection kicks in: 'According to C, whenever a fact grounds a principle, there's another principle which explains why that facts grounds the principle in question. You grant that the OP is dependent on facts. So, aren't the principles which underlie the OP your fundamental fact-insensitive principles of justice? Aren't you mistaking principles which are formulated in light of your fact-insensitive principles and facts, for your ultimate principles of justice?'

To respond to this challenge, the Rawlsian can grant that metaphysically fact-insensitive principles lie behind the OP. However, she can deny that those principles are principles of *distributive justice*. The Rawlsian adopts the OP in light of his (possibly) epistemically and metaphysically fact-insensitive principles of *pure procedural justice*, and the principle that people ought to be treated as equals. It has been observed by many commentators that some versions of utilitarianism, rights-based theories like Nozick's and egalitarian theories like Rawls's, offer competing interpretations of how people are to be treated as equals.[28] In this context, Rawls's constructivism can be in-

[28]Will Kymlicka, 'Rawls on Teleology and Deontology', *Philosophy and Public Affairs*, 17 (1988):3; Thomas Nagel, 'Equality', in: Andrew Williams and Matthey Clayton, edi-

terpreted as offering support for a controversial interpretation of equality through the device of the OP which relies on uncontroversial beliefs about procedural justice and the fundamental equality of human beings. In a letter to Nagel, quoted in Nagel's 'Equality', Rawls makes precisely this point:

> 'Suppose we distinguish between the equal treatment of persons and their (equal) right to be treated as equals. (Here persons are *moral* persons.) The *latter* is more basic: Suppose the Original Position represents the latter *re* moral persons when they agree on principles and suppose they *would* agree on *some* form of equal treatment. What more is needed?'[29]

This, then, is the structure of Rawls's theory. We have principles of procedural justice, and the principle 'People ought to be treated as equals' which are employed in the construction of the OP. These principles can be epistemically and metaphysically fact-insensitive. The denizens of the OP choose principles of distributive justice in light of facts. The principles of distributive justice which emerge may correspond to both our epistemically fact-insensitive and epistemically fact-sensitive principles. In cases where there are conflicts between our existing moral beliefs and the principles which emerge out of the OP, factual assumptions may be used to achieve a match. If factual assumptions are required to achieve this match, then our moral beliefs when we have reached reflective equilibrium will be an epistemically fact-sensitive set

tors, *The Ideal of Equality*, (Basingstoke: Palgrave MacMillan, 2002); Amartya Sen, *Inequality Reexamined*, (New York: Russell Sage Foundation, 1992). For an opinion to the contrary see Samuel Freeman, 'Utilitarianism, Deontology, and the Priority of Right', *Philosophy and Public Affairs*, 23 (1994):4.

[29]Rawls, quoted in Nagel, 'Equality', p. 79, emphasis in the original.

of beliefs.

Is the preceding account compatible with Rawls's own understanding of his theory? I have argued that Rawls does not deny C, and puts forward an altogether different thesis than not-C. The following passage from Rawls gives further support to this interpretation and shows that Rawls does not deny that moral principles, which may not be fact-sensitive, lie behind the OP. Rawls writes:

> '... [B]oth general facts as well as *moral conditions* are needed even in the argument for the first principles of justice... In a contract theory, these moral conditions take the form of a description of the initial contractual situation. It is also clear that there is a division of labor between general facts and moral conditions in arriving at conceptions of justice, and this division can be different from one theory to another.'[30]

Here, Rawls clearly accepts that moral beliefs are embodied in the OP and nothing he says suggests that those beliefs are held in light of facts, or that they are true only because certain facts obtain. This shows that Rawls, at least in *Theory*, did not hold the view Cohen attributes to him, and the theoretical structure I've attributed to Rawls is compatible with his own understanding of his theory.

Rawls also does not deny that we have beliefs about distributive justice prior to his theory. The goal of his theory is to systematize and give fur-

[30]Rawls, *Theory*, p. 138, emphasis added.

ther support to these beliefs, and revise some of them. As we saw, Rawls holds that the difference principle 'relies on the idea that in a competitive economy (with or without private property) with an open class system excessive inequalities will not be the rule'.[31] Commenting on this passage, Cohen argues that if Rawls 'appraised facts differently, he would reject the difference principle, *because it permitted too much inequality*'.[32] From this, Cohen concludes that there is a fact-insensitive background principle that Rawls is committed to, but has failed to articulate. It may be true that Rawls is committed to an epistemically fact-insensitive principle according to which too much inequality is objectionable. However, this doesn't commit him to the claim that too much inequality is objectionable in all possible worlds.

That we find some inequalities objectionable is quite explicit in Rawls's account. His goal is to show that his theory captures these beliefs. If his theory failed to capture this belief, then we would not have achieved reflective equilibrium. The belief set which includes the difference principle, and the belief that too much inequality is objectionable is an epistemically fact-sensitive set of beliefs. Let us suppose that this fact holds, and Rawls's theory is successful in capturing our moral beliefs in this world: Every inequality we find objectionable in this world is condemned by the difference principle. What should give way in a possible world in which huge inequalities are necessary to improve the condition of the worst-off? There are two different ways of treating this case. Firstly, we can take our intuitions about such worlds at face value. If we still find such inequalities troubling, then the theory has

[31] Rawls, *Theory*, p. 137.
[32] Cohen, *Rescuing Justice*, p. 259.

to give way. This is the approach which Rawls and utilitarians, as Rawls interprets them, object to. Alternatively we may refuse to take such intuitions at face value. We may query whether we would have the same intuition in a world in which huge inequalities were necessary, and how reliable our intuitions about worlds very different from ours are. This would amount to pointing out that the data by which the theory is to be tested, our intuitions in that possible world, is epistemically unavailable to us. This is the route that Rawls is in favour of. Note that this is a weaker claim than the one I attributed to Hooker in 5.3. It does not argue that the theory applies in all possible worlds, but merely shifts the burden of proof on those who claim that their intuitions about cases in all possible worlds are valid.

5.5 Conclusion

I have tried to show that in the passages Cohen uses to argue that Rawls denies C, Rawls does not deny C, but argues for a different claim. Rawls believes that we may legitimately use facts in deriving higher-level principles from lower-level ones, or hold fact-sensitive sets of principles. Along with this, Rawls believes that epistemically fact-sensitive principles are not necessarily metaphysically fact-insensitive principles. This allows Rawls to claim that principles of distributive justice are metaphysically fact-sensitive though they may be epistemically fact-insensitive.

I then examined the possibility that even though Rawls does not deny C, the truth of C poses difficulties for the OP. There were two possible

challenges. One was that because constructivist procedures rely on facts, the principles which emerge out of such procedures would have to be epistemically fact-sensitive principles. In response to this challenge, I showed how Kant's categorical imperative procedure could generate the duty that we should not make false promises, which is an epistemically fact-insensitive principle for most people. The second challenge was that Rawlsians are mistaken about the structure of their beliefs. What they take to be principles of justice are versions of their fundamental fact-insensitive principles which have been watered down with facts. In response to this challenge, I argued that the principles which underlie the OP are principles of procedural justice and the principle that people ought to be treated as equals, whereas the principles which emerge are principles of distributive justice. Therefore, the principles which emerge out of the OP are not the straightforward application of some principles to facts.

Part II

On the Design of the Original Position

Chapter 6

Uncertainty Behind the Veil of Ignorance

6.1 Introduction

In the previous chapters I focused on questions about justification and the nature of constructivism. I remained as non-committal as possible about the content of the correct principles of justice, and the design of the original position. In this chapter, having defended and clarified the nature of Rawls's constructivist project, I begin tackling its particular details, and address a challenge to Rawls's design of the original position. This challenge is best introduced by bringing in Harsanyi's employment of the veil of ignorance in his argument for utilitarianism.

In his article 'Cardinal Utility in Welfare Economics and in the Theory of Risk-taking', Harsanyi used a device similar to Rawls's veil of ignorance

to generate utilitarian conclusions.[1] Both Rawls and Harsanyi's veils of ignorance are intended to model the idea of impartiality. Behind both veils of ignorance, individuals deliberate on the assumption that they could end up in anyone's position. However, Rawls's veil of ignorance is in one respect thicker. Behind Rawls's 'thick' veil of ignorance, individuals do not know the likelihood that they may end up in any given person's position.[2] Behind Harsanyi's 'thin' veil of ignorance, they know that they have an equal chance of being in any person's position. This difference has substantial consequences. Under the thin veil of ignorance we have a decision problem under risk, in which case the maximization of expected utility is the preferred solution, and we end up with some form of utilitarianism. Under the thick veil of ignorance, we have a decision problem under uncertainty, in which case maximin, Rawls's preferred solution, is a plausible candidate as a solution.[3] Therefore, whether we use a thick or thin veil of ignorance is of great importance.

Rawls seems to lack a rationale for employing the thick veil of ignorance, other than the fact that it produces the results he prefers. This makes the

[1] John C. Harsanyi, 'Cardinal Utility in Welfare Economics and in the Theory of Risk-taking', *The Journal of Political Economy*, 61 (1953):5.

[2] There is another dimension in which Rawls's veil of ignorance differs from Harsanyi's, which I shall ignore. The denizens of the original position choose principles which will apply to groups, and not individuals. So what the denizens of Rawls's original position assume is that they can belong to any social group, but they lack probability estimates regarding the likelihood of their being a member of any given group.

[3] When we have probabilities for possible outcomes of our decisions, we face a *decision under risk*. When we don't have probability information, we face a *decision under uncertainty*. For instance if I know that an urn contains 8 red balls and 2 black balls, I can assign probabilities to the statements 'The ball I draw will be red' and 'The ball I draw will be black' coming out true, so my decision will be one under risk. If all I know is that some of the balls are red and some are black, then I can't assign probabilities to the previous two statements. In this case, my decision will be one under uncertainty.

whole procedure question begging. As Parfit puts it:

> 'Rawls himself points out that, since there are different contractualist formulas, he must defend his particular formula. This formula, he writes, must be the one that is 'philosophically most favoured', because it 'best expresses the conditions that are widely thought reasonable to impose on the choice of principles'. Could Rawls claim that, compared with the Equal Chance Formula [Harsanyi's formula], his No Knowledge Formula better expresses these conditions? The answer, I believe, is No. Rawls's veil of ignorance is intended to ensure that, in choosing principles, we would be impartial. To achieve this aim, Rawls need not tell us to suppose that we have no knowledge of the probabilities. If we supposed that we had an equal chance of being in anyone's position, that would make us just as impartial. Since there is no other difference between the Equal Chance and No Knowledge Formulas, Rawls's No Knowledge Formula cannot be claimed to be in itself more plausible.'[4]

The task of this chapter is to defend the characterization of the decision problem in the original position (hereafter the OP) as one of uncertainty rather than of risk. Before I present my defence, I'd like to set out its limits. Firstly, on some interpretations of probability, there is no difference between risk and uncertainty, because it is believed that all probability is subjective.

[4]Derek Parfit, *On What Matters*, ⟨URL: http://users.ox.ac.uk/~ball2568/parfit/parfit_-_on_what_matters.pdf⟩, Unpublished, 212-3. Nagel also makes the same objection. See Thomas Nagel, 'Rawls on Justice', *The Philosophical Review*, 82 April (1973):2, p. 11-12.

I'm here assuming that this is not the case. Secondly, I do not seek to defend the claim that maximin is the correct decision rule in the OP, or that the difference principle would emerge out of the OP. All I seek to establish is that the decision problem the denizens of the OP face should be characterized as one under uncertainty rather than risk.

There will be a lot going on in this chapter. Let me, therefore, present an intuitive road map for the reader. My central argument will be based on three claims. The first claim is that for some outcomes a, b, and c, where $a > b > c$, there isn't always a probability $\alpha, 0 < \alpha < 1$, such that $\alpha a + (1 - \alpha)c > b$. This is a denial of the continuity axiom in decision theory. My second claim will be that the denizens of the OP cannot precisely know for which outcomes continuity doesn't hold. My third claim will be that, given these first two claims, the decision problem in the OP should be interpreted as a decision problem under uncertainty, rather than risk, even if the denizens of the OP know that they have an equal chance of ending up in any given individual's position.

The bulk of this chapter is dedicated to defending the second claim. I defend this claim in the context of discussing the rejection of another claim about value, which is closely related to the rejection of the continuity axiom in decision theory: the Archimedean axiom. Suppose we have two objects of intrinsic value, a and b where $a > b$. According to the Archimedean axiom, for all such objects there is an amount k such that $kb > a$. In sections 6.2 and 6.3 I discuss cases from the domain of morality and prudential reasoning

where the violation of the Archimedean axiom is intuitively plausible. Following the recent literature, if for some a and b, where $a > b$, and there's no k such that $kb > a$, I'll say that a is superior to b. I offer definitions of the superiority relation and a brief account of the different ways in which superiority can come about in section 6.4. Even though the rejection of the Archimedean axiom is intuitively plausible, it is subject to a very forceful objection voiced by Griffin and Norcross, which I present in section 6.5. In section 6.6, I offer my diagnosis of their objections. I argue that their objection doesn't show what they intend it to show, namely that superiority ought to be rejected. Instead, it shows that we are not able to determine precisely where the superiority claims apply and don't apply when we're looking at the kind of cases they use in their examples. In section 6.7 I present my case for rejecting the continuity axiom of decision theory. In section 6.8 I tie up the different threads of my argument. I show that the decision problem in the OP should be characterized as a decision problem under uncertainty, because my diagnosis of what is going in Griffin and Norcross's examples is applicable here too.

6.2 Moral Cases

According to the Archimedean axiom, for any objects of intrinsic value, a and b, which are positively valuable and where a is more valuable than b, there is some amount k such that kb is more valuable than a. And in the case of objects of intrinsic disvalue, for any objects of intrinsic disvalue a and b where a is worse than b, there is an amount k such that kb is worse than

a. The Archimedean assumption governs much of the thinking about value, but there are cases where its rejection is intuitively very plausible. In this section, I shall discuss such cases from the domain of morality.

Suppose we face a choice between saving one person and saving five others. None of these people have any special claim to be helped. Neither are they different in other morally significant respects. We think that we should save the greater number. Views that aggregate the different interests of the people involved can accommodate this result with ease. Views that do not aggregate the interests of the people involved have so far failed to accommodate this result despite the plethora of ingenious attempts.[5]

Views which aggregate the interests of people are not without their problems. This example from Scanlon vividly illustrates the problem with them:

> 'Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over?'[6]

[5]For reviews of these attempts see Iwao Hirose, 'Review Article: Aggregation and Non-Utilitarian Moral Theories', *Journal of Moral Philosophy*, 4 July (2007):2; Michael Otsuka, 'Saving Lives, Moral Theory, and the Claims of Individuals', *Philosophy & Public Affairs*, 34 (2006):2.

[6]Thomas Scanlon, *What We Owe to Each Other*, (Cambridge, Mass: Belknap Press of Harvard, 1998), p. 235.

We think Jones should be rescued now, no matter how many people are watching the World Cup. We also think that the world in which Jones is rescued and the World Cup transmission is interrupted is better than the world in which Jones is not rescued and the World Cup transmission is not interrupted. Here, views that aggregate the different interests of the people involved fail to accommodate this result, or so it seems. Non-aggregationist views can accommodate this result.

One appealing solution for aggregationist views' failure to capture our intuitions in cases like Scanlon's transmitter room is to argue that the badness of Jones's pain cannot be outweighed by the badness of the frustration the interruption of the World Cup broadcast causes no matter how many people suffer from it. This means rejecting the Archimedean axiom.

It must be acknowledged that the rejection of the Archimedean axiom does not allay all worries about aggregationist views. Here are two cases where the revised aggregationist view, which rejects the Archimedean axiom, runs into problems, both of them due to Kamm. Suppose we can save from death either 1000 on one island or 1001 people who are on another island.[7] Kamm suggests that in this kind of case, the intuitively correct response is to toss a coin to decide what to do. The aggregationist view, even when it rejects the Archimedean axiom, would require saving the greater number. Another case Kamm offers is the following. We can either save A from death or we can

[7]F. M. Kamm, *Morality, Mortality: Death and Whom to Save From It*, Volume 1, (New York: Oxford University Press, 1993), p. 103.

save B from death and cure C's sore throat.[8] Kamm suggests, again, that the intuitively correct response is to toss a coin, because C's sore throat is an 'irrelevant utility'. The aggregationist view, even when it rejects continuity, would require saving B and curing C.

Insofar as one shares Kamm's intuitive responses, the aggregationist view, even when it rejects the Archimedean axiom, is not without problems.[9] However, we should keep in mind that theory choice is a comparative matter. If our choice is between (a) non-aggregationist views which cannot successfully justify saving the greater number, but give the *putatively* correct answers in Kamm's two cases and Scanlon's transmitter room case; (b) aggregationist views which justify saving the greater number, but give the *putatively* wrong answers in Kamm's two cases and Scanlon's transmitter room case; and (c) aggregationist views, which reject the Archimedean axiom, that justify saving the greater number and give the correct answer in Scanlon's transmitter room case, but give the *putatively* wrong answer in Kamm's two cases, then the last option is preferable.

Another argument against the Archimedean axiom can be found in Parfit's work in population axiology. Parfit asks us to consider the following cases.

[8]Kamm, *Morality, Mortality, Vol. 1*, p. 101-2.

[9]I must admit that I do not share Kamm's intuitions. However, these intuitions seem to be widely shared.

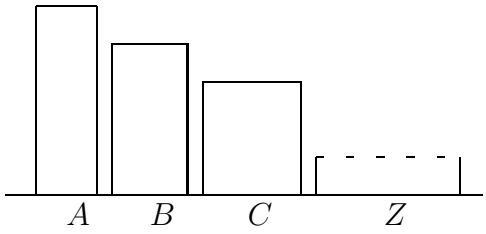

Figure 6.1: *The Repugnant Conclusion*

The height of each box represents the amount of well-being or utility enjoyed by each individual within that population, and the width of the boxes represents the number of people. Z 'is an enormous population all of whom have lives that are not much above the level where they would cease to be worth living'.[10] The problem with the principle purporting to maximize total utility is that for any possible population with a high quality of life, there is another larger population with lives barely worth living that the total utility principle recommends. Therefore, the total utility principle would recommend Z over A. Parfit calls this the Repugnant Conclusion.

As Parfit observes, rejecting the Archimedean axiom provides a way of avoiding the Repugnant Conclusion. When we move from A to Z, what Parfit calls the best things in life, 'the best kinds of creative activity and aesthetic experience, the best relationships between different people, and the other things which do most to make life worth living' are lost.[11]. By claiming that the Archimedean axiom doesn't hold for all of the goods under consideration, Parfit is able to block the inference from the total utility principle to the

[10]Derek Parfit, 'Overpopulation and the Quality of Life', in: Peter Singer, editor, *Applied Ethics*, (Oxford: Oxford University Press, 1986), p. 148.

[11]Parfit, 'Overpopulation and the Quality of Life', p. 161-2 Griffin also suggests this move. See James Griffin, *Well-Being: Its Meaning, Measurement and Moral Importance*, (Oxford: Clarendon Press, 1986), p. 339-40

Repugnant Conclusion. In Z the best things in life are lost, and this loss in value cannot be made up by the aggregate goodness of a stupendous amount of lives led in Z.

6.3 Prudential Cases

There are also cases in prudential reasoning where the rejection of the Archimedean axiom is intuitively plausible. Parfit presents one such case that is analogous to the argument for the Repugnant Conclusion. Suppose we face a choice between living a 100 years of life at extremely high quality and living forever a life in which the only good things in life are muzak and potatoes.[12] Parfit suggests that the first option is better.

We can also imagine a single person variant of Scanlon's example. Suppose you face a choice between suffering Jones's injury once, and suffering the disappointment of missing the World Cup countless times. (We can imagine that your life is much longer than the average human life in both cases, so that we are able to increase the number of times you are going to miss the World Cup.) Intuitively, the first option is worse.

A third example is due to Rawls. Rawls believes that once society reaches a certain level of development, his first principle, which ensures equal basic liberties, is lexically prior to his second principle that governs the distribution of wealth. After a certain point of development, basic liberties are not to be curtailed for gains in income and wealth no matter how large these gains are.

[12]Parfit, 'Overpopulation and the Quality of Life', p. 160.

In other words, Rawls believes that after a certain point the Archimedean axiom doesn't hold for the value of liberties and the value of income and wealth. Rawls's blanket talk of liberty, which has been forcefully criticized by Hart, makes Rawls's argument less persuasive then it would otherwise be.[13] When we consider specific liberties, his argument has force. For instance, it makes sense for an individual to be unwilling to trade off religious liberties for the sake of gains in income and wealth no matter how large these gains are.

6.4 Defining Superiority

The idea that there can be violations of the Archimedean axiom in values is often traced back to Mill. Mill is usually read as holding that some pleasures could be superior in quality in that no matter how much of the other pleasures there is, the superior pleasures will still be preferable.[14] Following Mill, and the recent literature, I shall refer to a good which can outweigh another no matter how much of the latter there is as a superior good.

One question which immediately arises about superiority is how it can come about. One possibility is that the superior good is of infinite value. If

[13]H. L. A Hart, 'Rawls on Liberty and Its Priority', *University of Chicago Law Review*, 40 (1973):3. Griffin who also notes that Rawls rejects the Archimedean axiom writes: 'The mistake here... seems to be to think that certain values—liberty, for instance—as types outrank other values—prosperity, for instance—as *types*. Since values, as types, can vary greatly in weight from token to token, it would be surprising to find this kind of discontinuity at the type—or at least at a fairly abstract type—level.'Griffin, *Well-Being*, p. 85-6, emphasis in the original.

[14]John Stuart Mill, 'Utilitarianism', in: *Collected Works*, Volume 10, (Toronto: University of Toronto Press, 1991), p. 211.

one of the goods in question is infinitely valuable and the other is not, then, no matter how much of the latter good we have, it will not equal the superior good in value. This is not the only way we can have superiority. If intrinsic value is not always additive, then, we can have cases where the value of each added unit of one good is diminishing, and no matter how much of that good we have, the total value will not go over a limit.[15]

I do not think that the proponent of superiority should commit himself to either of these accounts, or have a mathematical model of how superiority comes about. The rankings should come first, and the mathematical model(s) be built later. There may be different ways of modelling the same rankings and the proponent of superiority need not have any commitment other than his original – intuitively compelling – valuations.

We have been talking of any amount of a good being better than any amount of another. This is not the only possibility. Griffin, who is sceptical of the possibility that there are cases where 'any amount of A, no matter how small, is more valuable than any amount of B, no matter how large', suggests another possibility: 'enough of A outranks any amount of B'.[16] (Here A and B are objects possessing intrinsic value.) This gives us two different types of superiority:

> 'Strong superiority (roughly): Any amount of A is better than any amount of B.

[15]This possibility is raised in Wlodek Rabinowicz, 'Discussion - Ryberg's Doubts about Higher and Lower Pleasures - Put to Rest?' *Ethical Theory and Moral Practice*, 6 (2003):2.
[16]Griffin, *Well-Being*, pp. 83, 85.

> Weak superiority (roughly): Some amount of A is better than any amount of B.'[17]

Another possibility is what we can call conditional superiority. The value of certain goods which are superior to others can only be realized once someone has enough of other goods. For instance, intellectual satisfactions may be superior to the satisfaction of physiological needs. We may prefer a life of average length filled with intellectual satisfaction to any length of life where only physiological needs are satisfied. However, in order for intellectual needs to be satisfied, the satisfaction of physiological needs up to a certain level is a precondition. Conditional superiority may be combined with both strong and weak superiority. This gives us the following.

> Strong conditional superiority (roughly): If there is enough of B, then, any amount of A is better than any extra amount of B.
> Weak conditional superiority (roughly): If there is enough of B, then, some amount of A is better than any extra amount of B.

Conditional superiority provides one way of accounting for Rawls's idea that liberty is lexically prior once a certain level of development has been reached. The effective exercise of basic liberties is itself a value, or other superior values are realized through their effective exercise. This value is superior to the value of income. However, certain basic material needs have to be met for the effective exercise of these basic liberties.[18]

[17] Gustaf Arrhenius, 'Superiority in Value', *Philosophical Studies*, 123 (2005):1, p. 97.

[18] Saying that material needs are met, or physiological needs are satisfied, is in one way misleading. For it suggests that once these needs are satisfied we will have no more concern with them. This need not be so. For instance, our need for nutrition can be easily met with a basic diet, but we can pursue more pleasurable culinary experiences. In the case

6.5 Griffin-Norcross Sequences

So far we have looked at cases which suggest that the rejection of the Archimedean axiom is intuitively plausible, and examined different types of superiority. There are, however, certain cases where the commitment to superiority itself results in a contradiction. One such case is due to Alistair Norcross, who also observes that dropping the Archimedean axiom would bring aggregationist principles more in line with our intuitions.

Norcross, begins with an intuitively plausible superiority claim:

> '*Less*: The state of affairs constituted by any number of fairly minor headaches is less bad than even one premature death of an innocent person.'[19]

Norcross, then, proposes that we consider a sequence where at each step we decrease the gravity of the harm and increase the number of persons suffering from it. (Each step of the sequence, in effect, asserts that the Archimedean axiom holds for the two bads that figure in it.) The sequence takes this form:

1. 'one death is better than n^1 mutilations'

2. 'n^1 mutilations are better than n^2 xs (where x is some misfortune less bad than mutilation)'

.

of basic liberties and income, one could be interested in having more income than the amount necessary for the lexical priority of the liberty principle to kick in.

[19] Alastair Norcross, 'Comparing Harms: Headaches and Human Lives', *Philosophy and Public Affairs*, 26 (1997):2, p. 137.

.

.

n.$^{m-1}$ 'n^{m-2} broken ankles are better than n^{m-1} mild ankle sprains'

n.^{m} 'n^{m-1} mild ankle sprains are better than n^{m} mild headaches'.

We are inclined to assent to each of the claims in Norcross's sequence. Finally, by transitivity, we are required to accept that 'one death is better than n^{m} mild headaches', and reject Less.

Another similar sequence is offered by Griffin. Griffin first makes a superiority claim:

> 'Fifty years of life at a very high level of well-being—say, the level which makes possible satisfying personal relations, some understanding of what makes life worth while, appreciation of great beauty, the chance to accomplish something with one's life—outranks any number of years at the level just barely worth living—say, the level at which none of the former values are possible and one is left with just enough surplus of simple pleasure over pain to go on with it.'[20]

He then starts a sequence in one dimension of value:

1. 'Fifty years at a very high level—say, with enjoyment of a few of the very best Rembrandts, Vermeers, and de Hoochs—might be outranked

[20] Griffin, *Well-Being*, p. 86.

by fifty-five years at a slightly lower level-no Rembrandts, Vermeers, or de Hoochs, but the rest of the Dutch School.'

2. '. . . [F]ifty-five years at that level might be outranked by sixty years at a slightly lower level—no Dutch School but a lot more of the nineteenth-century revival of the Dutch School.'[21]

We repeat these steps until all appreciation of beauty is lost, and we are left with kitsch objects. We then repeat similar sequences for other dimensions of value Griffin mentions in his superiority claim, at which point we end up contradicting the initial superiority claim.

6.6 Explaining Griffin-Norcross Sequences[22]

In both Norcross's and Griffin's examples, we begin with an intuitively plausible superiority claim of the form: 'No amount of Z is better than any, or some, amount of A'. Nevertheless, we find ourselves willing to grant a claim of the form 'Some amount of B, which is slightly worse than A, is better than A.' Through similar steps involving gradual worsenings we arrive at the claim: 'Some amount of Z, which is slightly worse than Y, is better than Y.' Then, by transitivity we are required to concede that there is an amount of Z better than A, thereby contradicting our original superiority claim. We have three ways to resolve the inconsistency: (a) We can reject the supe-

[21] Griffin, *Well-Being*, p. 86.

[22] My argument in this section owes a great debt to Gustaf Arrhenius and Wlodek Rabinowicz, 'Value and Unacceptable Risk', *Economics and Philosophy*, 21 (2005):02, Mozaffar Qizilbash, 'Transitivity and Vagueness', *Economics and Philosophy*, 21 (2005):01, Richard Tuck, 'Is There a Free Rider Problem?' in: Ross Harrison, editor, *Rational Action: Studies in Philosophy and Social Science*, (New York: Cambridge University Press, 1979).

riority claim; (b) We can reject transitivity; (c) We can reject one of the comparisons we make at one of the steps of the sequence. Both superiority and transitivity are intuitively very plausible.[23] So, the best candidate for rejection is one of the comparisons we make.

The sequences presented by Norcross and Griffin work similarly to the sorites paradox. A sorites paradox takes us from a premise, which we take to be true, through an inductive premise, or repeated applications of a conditional, which we again hold to be true, to a conclusion, which we take to be false. The paradox of the heap, which gives the sorites paradox its name, has the premise '1 grain of sand does not make a heap', the inductive premise 'For any n, if n grains of sand don't make a heap then $n+1$ grains of sand don't make a heap', and the conclusion, depending on where we stop, '100,000 grains of sand do not make a heap'. The argument is valid, the premises seem true, but the conclusion seems false. The paradox arises due to the vagueness of the predicate 'heap'.

Vague predicates, like 'tall', 'red', 'heap', 'child', have borderline cases and lack sharp boundaries.[24] Borderline cases are cases where we can't decide whether or not a predicate applies. A related feature of vague predicates is their lack of well-defined extensions.[25] There isn't a sharp boundary which

[23]Temkin proposes rejecting transitivity in Larry Temkin, 'Worries about Continuity, Transitivity, Expected Utility Theory, and Practical Reasoning', in: Dan Egonsson et al., editors, *Exploring Practical Philosophy*, (Aldershot: Ashgate, 2001).

[24]Rosanna Keefe and Peter Smith, 'Introduction', in: Rosanna Keefe and Peter Smith, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997).

[25]Epistemicists deny this. They argue that vague predicates have well defined extensions, but their extensions are unknowable to us. For a defence of epistemicism see Timothy Williamson, *Vagueness*, (London: Routledge, 1994).

divides objects falling under a vague predicate and objects which don't fall under that predicate. For instance, we may not be able to decide whether someone is a child or not even though we know every observable fact about her. There isn't a determinate age beyond which someone is no longer a child. Small differences in the age of a person don't make a difference with regard to the appropriateness of calling them a child. Vague predicates' lack of sharp boundaries ensures that we assent to the conditional premise of the Sorites paradox.

This is, roughly, what I take to be going on in Norcross-Griffin sequences. The goodness or badness of some outcomes cannot be directly perceived, or judged, by us. The badness or goodness of all outcomes, and how they compare to other outcomes, cannot always be read off just by being presented with them. We need to rely on re-descriptions of outcomes which reduce them to more basic judgements. This process of re-description can be carried out consciously or unconsciously. These redescriptions contain vague predicates. Accordingly, we are unable to identify the specific points where the predicates in our re-descriptions no longer apply. As a result, we feel compelled to assent to the trade-offs in the consecutive steps of Griffin-Norcross sequences.

The redescription process can take place in two different ways. I think both processes actually take place, but either of them taking place is enough for my argument. On the first view, we think about the options we are comparing by consciously re-describing them in natural language. We do this in order to make the options vivid and establish connections between the outcomes

under consideration and our more basic judgements. This is supported by attending to the way we make actual comparisons. When I think about the comparisons in Norcross' sequence, I find myself formulating alternative descriptions of the options like the following: 'Very painful', 'Unbearably painful', 'Disabling', 'Impossible to ignore', 'Easy to ignore', 'One could carry on as usual despite it', 'For the length of the average lifetime', 'For a very short time' etc.[26]

On the second view, the re-descriptions do not take place in natural language, but in an innate language of thought, sometimes called *mentalese*, that functions as the medium of thought.[27] Mentalese is like a natural language in that it uses words like natural languages and these words combine to form mental sentences. On this view, the redescriptions process can be done consciously or unconsciously. Since the language of thought has to be translatable into natural languages – otherwise we wouldn't be able to express our thoughts or understand the thoughts of others – it has to contain vague predicates.[28]

On both views, we do not need to identify a single term that should bear the blame for the introduction of vagueness into our re-descriptions. It is unlikely that we rely on a single term when trying to make the alternatives

[26]This is basically an extension of Qizilbash's account. See Qizilbash, 'Transitivity and Vagueness', p. 119-120.

[27]The *locus classicus* is Jerry A. Fodor, *The Language of Thought*, (Cambridge, Mass: Harvard University Press, 1975).

[28]For the argument that the language of thought has to be vague see Roy A. Sorensen, 'Vagueness within the Language of Thought', *The Philosophical Quarterly*, 41 October (1991):165.

we are comparing vivid. What is more likely is that there are many terms we use in describing the alternatives which are susceptible to borderline cases.

It is useful to explicitly state the premises of our argument. Our first premise is that when judging whether an option is better than, or superior to, another, we need to represent it mentally. The second premise is that this medium of representation is a language. This language may be a natural language or may be mentalese. The third premise is that this language, which acts as the medium of thought, contains vague predicates.[29] The fourth premise is that some of our value judgements are derived from more basic value judgements. Sometimes when deciding whether A is better than, or superior to, B our judgement will be inferential. It will have, roughly, the form A possesses properties a, b, $c\ldots$, B lacks these properties, hence A is better than B. I do not claim that this process of reasoning needs to be consciously carried out.

Let me now illustrate my account using Griffin's example as our basis. For ease of exposition, let us assume that the only relevant value is the enjoyment of beauty. Then, the basic superiority claim is that 'Fifty years of life spent in enjoyment of the arts is better than any duration of life spent in enjoyment of kitsch objects'. On my account, to evaluate the superiority claim and to make the comparisons in Griffin's sequence, we need to make it vivid and spell out both 'enjoyment of art' and 'enjoyment of kitsch'. Let's say the predicates n_1, n_2,$\ldots \mathrm{n}_m$ are the ones we use in our account of 'enjoyment of

[29]The second premise might be replaced by a weaker claim. What matters for our purposes is that the medium of representation contains vague predicates.

art'. Some of these terms will be used in our specification of what amounts to art, and some will be used in our specification of what amounts to enjoyment. Some of these predicates will be vague and lack sharp boundaries. As a result, we won't be able to determine the precise cut-off point when an object is no longer a work of art, or when our relationship with a work of art can't be characterized as enjoyment.

Description	n_m	$\neg n_m$
	.	.
	.	.
	.	.
	n_3	$\neg n_3$
	n_2	$\neg n_2$
	n_1	$\neg n_1$
Value Claim	Beauty	Kitsch

Table 6.1: *Vague predicates and the superiority claim*

The diagram above illustrates our predicament. The dotted vertical line signifies the absence of sharp boundaries. This means that we are unable to tell at which specific point $n_1 \ldots \mathrm{n}_m$ no longer apply to objects. The straight vertical line signifies the presence of sharply defined cut-off point. We are committed to the superiority of a life spent in enjoyment of the arts to a life spent in enjoyment of only kitsch objects. We are committed to the claim that a life that $n_1 \ldots \mathrm{n}_m$ is applicable to is superior to one that they aren't

applicable to. We are also committed to making trade-offs between quality and quantity as long as we are in the domain of beauty. At each step of Griffin-Norcross sequences, we ask whether $n_1 \ldots n_m$ do apply to the objects under consideration. If they do, then we are willing to make trade-offs. Since, $n_1 \ldots n_m$ contain vague predicates, there isn't a sharp cut-off point where we can say $n_1 \ldots n_m$ no longer apply, even though $n_1 \ldots n_m$ clearly apply to some objects, and clearly don't apply to others. Since our fundamental superiority claim is that the objects $n_1 \ldots n_m$ apply to are superior to those $n_1 \ldots n_m$ don't apply to, we can't determine the precise point where we should stop making trade-offs. For this reason, we find ourselves accepting the trade-offs offered at each consecutive step of Griffin-Norcross sequences.

The role of redescriptions in making the comparisons in Norcross-Griffin sequences confirms a point made by Dorsey in his insightful discussion of Norcross's sequence. Dorsey claims that 'the fulfilment of deliberative projects is strongly superior to hedonic values, though both are important for a complete account of human welfare'.[30] He then offers the following diagnosis of Norcross's sequence. Norcross's sequence goes wrong in focusing only on one dimension of value: the intrinsic badness of physical discomfort *qua* physical discomfort. This is a mistake, because physical discomforts at certain levels 'become instrumentally injurious to our deliberative projects'.[31] So, it is at the point where deliberative projects, which are superior to hedonic values, become impossible to carry out that the Archimedean axiom doesn't hold.[32]

[30]Dale Dorsey, 'Headaches, Lives and Value', *Utilitas*, 21 (2009):01, p. 46.

[31]Dorsey, 'Headaches', p. 49.

[32]Note that on Dorsey's account, deliberative projects are what we called conditional superior goods.

This account, and the discontinuity claim Dorsey offers are I believe correct. (In fact, I shall rely on a similar claim in the next section.) And it is confirmed by attending to how we redescribe the options in Norcross's sequence. Recall some of the descriptions we used in our descriptions: 'Unbearably painful', 'Disabling', 'Impossible to ignore', 'Easy to ignore', 'One could carry on as usual despite it'. They point to how the physical pain relates to one's life projects. So, the redescription process not only helps us get a handle on the magnitude of pains, but also directs our attention to the place of pains within a wider network of values. On Dorsey's proposal, we have a sharp cut off point – the point at which physical pain becomes injurious to our deliberative projects. However, if our account of the Norcross-Griffin sequences is correct, where that point is can't be specified with precision due to the vagueness of the terms we rely on to spell out life projects, physical discomforts, and the relationship between them.[33]

6.7 The Rejection of Continuity

The denizens of the OP are deliberating in a probabilistic framework. They are not ranking outcomes that are certain. Instead they are considering possible outcomes with differing likelihoods. Accordingly, the rejection of the Archimedean axiom that we looked at in the previous sections is not directly relevant to the decision problem in the OP. However, there's a related thesis

[33]I should point out that Dorsey's proposal addresses a different puzzle than the one I'm addressing. I'm addressing the puzzle about why we would assent to the claims in each step of the Norcross-Griffin sequences. Dorsey is addressing a puzzle about how superiority can arise when the only aspects of the objects under consideration that are relevant to their value can be placed on a continuum.

that is relevant to the deliberations of its denizens: the continuity axiom of decision theory. According to the continuity axiom, for all outcomes a, b, and c, where $a > b > c$, there is a probability $\alpha, 0 < \alpha < 1$, such that $\alpha a + (1 - \alpha)c > b$.[34]

Even though continuity has often been defended as a technical requirement, it has substantive consequences and has not gone unchallenged.[35] There are various counterexamples to the continuity axiom offered in the decision theory literature.[36] The following example is due to Kreps.[37] Suppose there are three outcomes. Outcome a gives you \$1000; outcome b gives you \$10; and in outcome c you are killed. Continuity requires that there is a probability, p, $0 < p < 1$, such that $pa + (1 - p)c > b$. That is, one ought to be willing to run a small risk of getting killed to end up with \$1000 instead of \$10. This is highly counter-intuitive.

[34]See John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, 3rd edition. (Princeton, N.J.: Princeton University Press, 1953), p. 26 for a presentation of this axiom.

[35]The rejection of continuity means that the relevant value function cannot be represented with a real-valued function. Ulrich Schmidt, *Axiomatic utility theory under risk: Non-Archimedean Representations and Application to Insurance Economics*, (Berlin: Springer, 1998), p. 69. It doesn't cause any further problems. Preferences that violate continuity, but respect all other axioms of Von Neumann and Morgenstern, can still be represented by a utility function, and the notion of rationality as maximization remains intact. Peter C. Fishburn, *Nonlinear Preference and Utility Theory*, (Brighton: Wheatsheaf, 1988), p. 47.

[36]Armen A. Alchian, 'The Meaning of Utility Measurement', *The American Economic Review*, 43 March (1953):1, p. 36-7; John S. Chipman, 'The Foundations of Utility', *Econometrica*, 28 April (1960):2, p. 221; R. Duncan Luce and Howard Raiffa, *Games and Decisions: Introduction and Critical Survey*, (New York: Dover, 1989), p. 27.

[37]David M. Kreps, *Notes on the Theory of Choice*, (Boulder, Colo: Westview, 1988), p. 45.

There's also a standard way of dealing with such objections. Kreps offers it right after presenting the previous putative counter-example to continuity.

> 'Suppose I told you that you could either have $10 right now, or, if you were willing to drive five miles (pick some location five miles away from where you are), an envelope with $1000 was waiting for you. Most people would get out their car keys at such a prospect, even though driving the five miles increases ever so slightly the chances of a fatal accident. So perhaps the axiom isn't so bad normatively as may seem at first sight.'[38]

We can make two preliminary responses. Firstly, we should observe that Kreps's second example shows only that it might not be irrational to take the gamble that was offered in Kreps's first example. It doesn't establish the stronger claim that continuity is required by rationality. Secondly, we ought to take it as a warning that we should not be too hasty in judging that continuity doesn't hold between certain outcomes. All that the example shows, if it shows anything, is that the rejection of continuity in Kreps's first example was mistaken.

Nevertheless, there's something puzzling. What explains the divergence of our intuitions in these two cases? Kreps's second example transforms the first example into the kind of choice we face in everyday life. We of course need to do several things that might increase the chance of us dying to pursue our goals, make a living etc. Our previous discussion of conditional superiority may help us here. Suppose that continuity between life and money is to be

[38]Kreps, *Notes*, p. 45-6.

rejected conditionally. It is only if you've guaranteed a certain level of income and wealth that continuity breaks down. After all we want to live longer for the values life embodies. For the enjoyment of these values – and more fundamentally, for sustenance – a certain level of wealth is a prerequisite. The rejection of continuity between life, extra income, and death that motivates our response to Kreps's first example should be interpreted as the rejection of continuity between a certain kind of life – one that embodies certain values which are only possible given some level of income and some amount of risk taking –, extra income, and death. So, for those who are committed to the conditional rejection of continuity between life and money, one *does* need to take risks which will increase the risk of death when those risks are necessary for securing the conditions that give life its value. It's not the duration of time one spends alive but the goods a life embodies that matter. Cases like Kreps's second example put one in the mindset of such choices where we need to take risks to secure the goods that are a precondition for a good life. One thinks, 'Well of course, I drive to work everyday. The choice in Kreps's example is no different from driving to work. So I should take the risk'. However, Kreps's first example does not put us into that mindset. Therefore, Kreps's second example should be taken as a sign that the rejection of continuity in Kreps's example is conditional.

What is the discontinuity claim relevant to the OP? The denizens of the OP are examining principles that will determine their long-term expectations of primary goods. Each social position that results from the application of a principle they choose entails, in effect, a certain life prospect. Overall life

prospects are cases where the possibility of discontinuity is highly plausible.[39] That is, it is more plausible to expect discontinuity between lives taken as a whole rather than between experiences of shorter duration. The basic, and admittedly very rough, idea is that continuity doesn't hold when one option is a life in which S is guaranteed the ability to do something with his life, and the other option offers either (a) such a life plus some further goods, or (b) a life in which S cannot do something with his life, even if this latter possibility is very unlikely.

It is carrying out plans of life and projects we think worth pursuing and forming deep relations that give life meaning. A life containing these values is to be preferred to a life where they are absent, even though a life in which they are largely absent may, nevertheless, be one worth living. Rawls's primary goods are instrumental to us carrying out our plans of life, and with more primary goods we can pursue our plans and projects with greater ease, or pursue more ambitious plans and projects. However, the possibility of having more primary goods doesn't justify taking gambles where the other possibility is losing the capacity to pursue plans and projects, and to form deep relations.

The account of how vague predicates figure in our evaluation of different alternatives that we presented in section 6.6 is applicable to the decision problem in the OP, because the outcomes under consideration are very complex and we can get a handle on them only by re-describing them. There are

[39] I use 'discontinuity' to refer to cases where the continuity axiom doesn't hold. This should not be confused with Griffin's use of 'discontinuity' in his discussion of superiority.

many points at which vagueness can creep into these descriptions. The idea of doing something with one's life itself is vague. And even if we could specify it in greater detail, our specification would contain many vague terms. Accordingly, if discontinuity sets in for the decision problem in the OP, then we will be unable to precisely identify the point where it sets in. The problem isn't identifying the precise amount of primary goods necessary for pursuing plans and projects that give life meaning. We are assuming that denizens of the OP know all the relevant empirical facts. The problem is identifying the precise point after which a person can be said to be not pursuing plans and projects that give a life meaning.

6.8 Applying the Account to the OP

Let us, finally, apply the account of discontinuity we have developed to the decision problem in the OP. We begin by noting an ambiguity in discussions of the decision problem in the OP. The decision problem in the OP is often discussed as if primary goods were a theoretical construct analogous to utility. Maximinning one's expectation of primary goods and maximinning one's expectation of utility is treated as being equivalent. This way of looking at the decision problem in the OP brings into sharp relief the difference between Rawls's principles and utilitarianism. However, this assumed equivalence is misleading. Rawls, in fact, attributes a certain utility function to the denizens of the OP, which determines their valuation of primary goods, and relies on this function in his argument for his principles of justice.[40] (I shall refer

[40]Rawls concedes that this is the case. See John Rawls, *Justice as Fairness: A Restatement*, (Cambridge, Mass: Harvard University Press, 2001), p. 107.

to this function as a value function to distinguish it from an individuals' subjective utility functions).

These are some of the ways in which Rawls specifies the value function of the denizens of the OP. Its denizens know that 'they normally prefer more primary social goods rather than less'.[41] They value the interests protected by Rawls's first principle more than they value those promoted by his second principle.[42] A person in the OP has 'a conception of the good such that he cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact, be sure of by following maximin' and 'the rejected alternatives have outcomes that one can hardly accept'.[43]

Once we acknowledge that the denizens of the OP have a value function, we can distinguish between two different characterizations of the decision problem they face. The decision problem faced by the denizens of the OP can, initially, be characterized in this way:

		Individual positions				
		i_1	i_2	.	.	i_n
Principles	A_1	p_{11}	p_{12}	.	.	p_{1n}
	A_2	p_{21}	p_{22}	.	.	p_{2n}
	.	.	.	.	.	.
	.	.	.	.	.	.
	A_n	p_{n1}	p_{n2}	.	.	p_{nn}

Table 6.2: *The primary goods characterization of the decision problem*

[41] John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 123.
[42] Rawls, *Theory*, p. 131.
[43] Rawls, *Theory*, p. 134.

An outcome p, which is given in terms of primary goods, corresponds to each pair consisting of a principle of distribution and individual position. Since there's a value function which determines the value of each outcome for the denizens of the OP, the decision problem needs another step, which transforms outcomes in terms of primary goods to values. Once the values are known, the decision problem will be the following:

		Individual positions				
		i_1	i_2	.	.	i_n
Principles	A_1	v_{11}	v_{12}	.	.	v_{1n}
	A_2	v_{21}	v_{22}	.	.	v_{2n}
	.	.	.	.	.	.
	.	.	.	.	.	.
	A_n	v_{n1}	v_{n2}	.	.	v_{nn}

Table 6.3: *The value characterization of the decision problem*

Once this intermediate step becomes apparent, the question of the nature of the value function that takes us from primary goods to value arises. Given that we are deciding on the value function of the denizens of the OP, we can specify the function in such a way that maximinning one's expectation of primary goods is the correct decision rule even under conditions of risk. That is, we can specify the value function in such a way that the outcome which maximizes expected value is the one which maximins primary goods. Rawls notes that this would be the case if the value function has a sharp bend at the point guaranteed by maximin.[44] Of course, specifying the value function in this way would make the OP effectively redundant.

[44]For Rawls's discussion of this point and a graph illustrating this possibility see Rawls, *Justice as Fairness*, p. 107-9. There are a few other steps to his argument which I've ignored here.

What we need is a way of specifying the value function of denizens of the OP that is acceptable on its own, and doesn't make the argument from the OP trivial. My argument has been that (i) continuity must be rejected, and (ii) we, and *a fortiori* the denizens of the OP, cannot know where discontinuity sets in. If this argument is correct, then there is a way of specifying the value function which is acceptable on its own, and justifies the imputation of uncertainty rather than risk to the OP. It would follow that, under certain further assumptions, the constructivist procedure producing the difference principle could be justified.

Suppose the denizens of OP know that they have an equal chance of ending up at any individual position. Also assume that they know their society's level of economic development.[45] This does not suffice for them to know which principles are preferable. To know this, they need to consult the value function. As I've argued, the axiom of continuity in the value function must be abandoned. The rejection of continuity requires that the denizens of OP should reject certain lotteries. Furthermore, we don't know where precisely the relevant discontinuity in that function to sets in. There is, therefore, an uncertainty as to which lotteries the denizens of OP should reject. Unlike Rawls's OP, in this case, the uncertainty is due to the nature of the value function and our inability to determine where precisely discontinuity sets in. The denizens of the OP know that they have an equal chance of ending up

[45]This assumption is in conflict with Rawls's characterization of the veil of ignorance. I make this assumption to emphasize the robustness of the argument we're developing here. Not knowing a society's level of development adds an element of uncertainty. I want to show that the decision problem in the OP can be characterized as a decision problem under uncertainty even without this assumption.

as any individual, but they don't know exactly how each outcome maps on to each value (the mapping of each p_{ij} to each v_{ij}). More precisely, they know that there are certain outcomes which they should avoid, lotteries they should reject, but they don't precisely know which outcomes and lotteries these are.

The situation of the denizens of the OP can be likened to the following. Suppose you are ill and unless you receive treatment you will die. There is only one doctor who can cure you. She keeps her fee a secret. You only know that it's an amount between £110 and £0. There are three lotteries you can choose from. Each lottery has three possible outcomes all of which are equally likely. The money you may win from these lotteries is your only source of money for covering the doctor's fee. The first lottery has as its prizes £60, £50, and £40. The second lottery has as its prizes £80, £80, and £10. The third lottery has as its prizes £110, £30, and £10. Which lottery should you choose?

Viewed in one way, in this example, there's no uncertainty about the outcomes. You are in a position to calculate the expected values, in money terms, of the lotteries. However, there's uncertainty as to which lottery gives you the greatest chance of receiving treatment the treatment you need, which is your primary concern. If for instance, the doctor's fee is between £40 and £10, then the first lottery gives you a greatest chance. If the doctor's fee is £100, or more, the third lottery gives you the greatest chance.

In this example, you have probability estimates for the likelihood of ending up with any given amount of money. So, if we focus on money, you face a decision problem under risk. Similarly, the denizens of the OP, who know they have an equal chance of being anyone, don't face uncertainty regarding the outcomes in terms of primary good – they know the probability of ending up with any given amount of primary goods. There's, nonetheless, uncertainty about the value of the outcome. In both cases, there's a threshold that the decision maker would like to pass, but doesn't know where that threshold is. Hence the uncertainty.

Having presented the intuitive argument for the characterization of the decision problem in the OP as one under uncertainty – despite the availability of probability estimates – we can present a more formal account. Let us suppose that the denizens of the OP are behind a thin veil of ignorance like Harsanyi's. They know the probability of ending in any single position. Suppose there are two principles of distribution, A and B, being considered by the denizens of the OP. Let $\mathrm{o}_1, \ldots, \mathrm{o}_n$ denote different amounts of primary goods individuals can end up with, where every element in the sequence contains progressively lower amount of primary goods such that $o_1 > o_2 > \ldots > o_n$.

Suppose we think that continuity doesn't hold between o_2 and o_n. There is no probability $0 < \alpha < 1$ such that $\alpha o_1 + (1 - \alpha)o_n > o_2$. That is, gains above o_2 do not justify taking gambles where one of the outcomes is o_n no matter how small the likelihood of ending up with o_n. Let's say we

are considering two principles of distribution. One of them guarantees o_2. Under the other principle we may end up with either o_1 or o_n. Given the existence of discontinuity between o_2 and o_n, the denizens of the OP should pick the first principle no matter how small the probability of ending up with o_n under the second principle is.

Since we can't determine precisely where discontinuity sets in, we can construct a Griffin-Norcross sequence that will lead us to contradict this discontinuity claim by making us agree to trade-offs on consecutive outcomes on the sequence. Suppose we are asked whether there's a probability $0 < \alpha < 1$ such that $\alpha o_1 + (1 - \alpha)o_3 > o_2$. Our response is that there is such a probability. We then repeat this procedure for each consecutive outcome, and end up contradicting our initial discontinuity claim. If our account of the real lesson of the Griffin-Norcross sequence is correct, there is a point where one of these comparisons is wrong, but we don't know which. Let's call the outcome where discontinuity actually sets in o_c. Let p_{ai}, $1 > p_{ai} > 0$, represent the probability of ending up with some outcome, o_i, under principle A. Let p_{bi}, $1 > p_{bi} > 0$, represent the probability of ending up with any given outcome, o_i, under principle B.[46] Since we don't know where o_c is we don't know the values of $\sum_{i=1}^{c} p_{ai}$, $\sum_{i=c}^{n} p_{ai}$, $\sum_{i=1}^{c} p_{bi}$, and $\sum_{i=c}^{n} p_{bi}$ even though we know the values of each p_{ai} and p_{bi}. In other words, we don't know the probability of ending up above or below the point where discontinuity sets in under these two principles even though we know the probability of ending up with

[46]Since we're assuming that the denizens of the OP have an equal chance of ending up as any given individual, the probability of ending up with any given amount of primary goods is given by the number of people in a society who have that amount of primary goods over the total number of people in society.

any given share of primary goods under both principles. Accordingly, their decision problem is one under uncertainty.

6.9 A Quick Solution Rejected and a Suggestion

In this section, I'd like to consider a possible solution to the problem I've been examining and show its weakness. The basic idea I've been labouring is this:

1. Continuity doesn't hold when one option is a life in which S is guaranteed the ability to do something with his life, and the other option offers either (a) such a life plus some further goods, or (b) a life in which S cannot do something with his life, even if this latter possibility is very unlikely.

2. This claim should be unpacked as: Continuity doesn't hold when one option is a life which has the set of properties X, and the other option offers either (a) such a life plus some further goods, or (b) a life which lacks the set of properties X, even if this latter possibility is very unlikely.

3. Given the vagueness of our concepts we are unable to tell the exact cut-off point where a life no longer possesses X.

4. As a result, we are unable to tell where discontinuity sets in.

5. The same point applies to the denizens of the OP. Denizens of the OP know that continuity doesn't hold between certain outcomes. However they don't know where the cut-off point is. For this reason their decision can be characterized as one under conditions of uncertainty even though they have all the relevant information that we can provide them.

The proposed solution I'd like to consider, which takes its cue from supervaluationist accounts of vagueness, is as follows.[47] Our concepts may be vague, but there are cases where we are able to give a definite answer as to whether an object possesses a certain property. For instance, we are able to tell (definitely) that a single grain of sand is not a heap. We can classify certain collections of sand grains as being definitely heaps, and some as being definitely not heaps. We can do the same for the outcomes faced by the denizens of the OP. Some outcomes are definitely above the threshold where discontinuity sets in and some are definitely below it. The denizens of the OP, behind the thin veil of ignorance, should set a guaranteed minimum at the point with the least amount of primary goods they think is definitely above the threshold. They should then treat the decision problem for outcomes above it as one under risk.[48]

This solution is, however, untenable due to the existence of higher-order vagueness. We've seen that vague predicates lack sharp boundaries and admit

[47]For formulations and defences of supervaluationism see Rosanna Keefe, *Theories of Vagueness*, (Cambridge: Cambridge University Press, 2000) and Kit Fine, 'Vagueness, Truth and Logic', in: Rosanna Keefe and Peter Smith, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997).

[48]In effect, they'd be choosing 'the principle of average utility constrained by a certain social minimum'. Rawls, *Theory*, p. 278.

of borderline cases. Similarly, 'cases that are borderline' lack sharp boundaries: 'The limits of vagueness themselves are vague'.[49] For instance, the predicate 'child' is vague. There are some cases where it's not clear whether someone is a child or not. There isn't an n^{th} second of a person's life after which we'll say 'He's no longer a child'. Similarly, 'definitely a child' is vague. There isn't an n^{th} second in a person's life after which we'll say 'He is no longer definitely a child'. The same point applies in the OP. Even though there will be outcomes which we think are definitely above the threshold, we will not be able to specify the exact point above which outcomes are definitely above the threshold.[50]

This proposed solution, however, suggests one way in which we can enrich the information available to the denizens of the OP. Even though they do not know where discontinuity sets in, they may have a rough estimate. For instance, they may know that o_c, the point discontinuity sets in, is closer to o_n than it is to o_1, or that it is closer to o_1 than it is to o_n. Information like this can offer considerations in favour of specific decision rules such as maximin.

[49]Williamson, *Vagueness*, p. 2. Russell puts the same point more elaborately: 'The fact is that all words are attributable without doubt over a certain area, but become questionable within a penumbra, outside which they are again certainly not attributable. Someone might seek to obtain precision in the use of words by saying that no word is to be applied in the penumbra, but unfortunately the penumbra itself is not accurately definable, and all the vaguenesses which apply to the primary use of words apply also when we try to fix a limit on their indubitable applicability.' Bertrand Russell, 'Vagueness', in: Rosanna Keefe and Peter Smith, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997), p. 63-4.

[50]I'd like to emphasize that I'm not arguing in favor of maximin, or any other decision rule. So this response isn't intended to support a specific decision rule. Its aim is only to defend the characterization of the decision problem in the OP as one under uncertainty rather than risk.

6.10 Conclusion

By way of conclusion, I'd like to mention one implication of the argument offered here, which I find appealing. According to the account offered here the decision problem in the OP should be characterized as one of uncertainty, because discontinuity holds for the outcomes the denizens of the OP face and they cannot know precisely where discontinuity sets in. This makes the characterization of the decision problem in the OP, and the principles of distribution chosen in the OP, sensitive to the outcomes under consideration. For instance, if the denizens of the OP were facing outcomes in which continuity holds, the decision problem would be one under conditions of risk. In cases where discontinuity holds, the decision of the denizens of the OP will be sensitive to their rough estimate as to where discontinuity sets in. If they believe discontinuity sets in at a point closer to the worst outcomes, the principles they choose will reflect this. Because it is sensitive to the outcomes under consideration, the OP could be used to accommodate different principles of distribution. Given the variety in our intuitions about which distributions are acceptable, and the failure of a single distributive principle to capture our intuitions, this is a welcome result worth exploring.

Part III

Cohen's Egalitarian Ethos

Chapter 7

Egalitarian Ethi and Primary Goods

7.1 Introduction

The first part of this thesis was an examination of Rawls's constructivism and a response to Cohen's metaethical challenge to it. In this and the next chapter, I turn to one of Cohen's normative challenges to Rawls's theory of justice: his argument that justice demands an egalitarian ethos. As I indicated in chapter 4, I believe that, in the absence of a concern with publicity and practicality, welfare remains a plausible candidate for the metric of justice. Accordingly, I shall evaluate the case for an egalitarian ethos without committing myself to a specific account of the metric of justice. In this chapter, I shall assume that primary goods are the metric of justice, and evaluate the case for an egalitarian ethos with that assumption in place. In the next chapter, I shall assume that welfare is the metric of justice, and and evaluate

the case for an egalitarian ethos on that assumption.

I don't wish to deny the desirability of an egalitarian ethos when it is broadly construed. I'm convinced that several attitudes and everyday choices should be governed by an egalitarian ethos in a just society. My target in these two chapters is an egalitarian ethos narrowly construed. By an egalitarian ethos, I shall understand a moral duty which specifies how one should act in the market, particularly when deciding how much to work and in which occupations to work, aimed to make the co-realization of equality and Pareto-efficiency possible. Accordingly, proposals for an egalitarian ethos which tell us how we should interact with each other are not the subject of my criticism. Neither are arguments which hold that we should not take up certain occupations as a matter of justice when this requirement is not based on an effort to reconcile equality and Pareto-optimality. For instance, an egalitarian ethos which required that individuals refrain from taking up movie roles that sustain racial stereotypes, even if a legal ban on such movie roles does not, and should not, exist, would not be a target of my criticism. My goal in these two chapters will be to establish that the end that the egalitarian ethos is offered by Cohen as a means for is undesirable, or that the egalitarian ethos is unnecessary or unable to achieve these ends.

Let us approach the case for an egalitarian ethos by looking at Brian Barry's reconstruction of John Rawls's informal argument for the difference principle, which has been dubbed the Pareto argument for inequality by

Cohen.[1] The argument has two stages. The first stage of the argument seeks to establish the *prima facie* justice of an equal distribution. The basic idea is that equality of opportunity requires the elimination of all arbitrary causes of inequality, and that all causes of inequality, such as one's social background and one's genetic endowments, are morally arbitrary. Therefore, equality of opportunity requires equality of outcome.[2] The second stage of the argument, which is the subject of this chapter, purports to show why certain inequalities are justified. We begin with an equal distribution of primary goods, D1.[3] It is then argued that certain inequalities in the distribution of primary goods, which act as incentives to the talented, can result in benefits to the worst-off.[4] Thus we should move from D1 to this unequal distribution, D2, in which everyone has more primary goods than they do in D1.[5]

In his powerful critique of the Pareto argument for inequality, Cohen points out that when D2 is feasible, another distribution, D3, which is equal and Pareto-efficient, is also feasible.[6] Since D3, an equal and Pareto-efficient distribution, is feasible the Pareto principle does not justify departures from

[1]Brian Barry, *Theories of Justice: A Treatise on Social Justice, Vol. 1*, (Berkeley: University of California Press, 1989), p. 213-235.

[2]Barry, *Theories of Justice*, p. 224. I do not think this a correct reading of Rawls's argument in section 12 of *A Theory of Justice*. However, I shall set that issue aside. What matters for our purpose is that at the end of the first stage of this argument, there is a presumption in favour of an equal distribution of primary goods, which is something I do not contest.

[3]It is important to my argument that it is equality of *primary goods* and not some other metric that is deemed *prima facie* just in the first stage of the Pareto argument for equality. On this I diverge from Cohen's treatment of the argument. See G. A Cohen, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008), p. 94, n. 21.

[4]Following Cohen, by talented, I mean those 'who command a high salary in a market economy'. See Cohen, *Rescuing Justice*, p. 119.

[5]Barry, *Theories of Justice*, p. 231-234.

[6]Cohen, *Rescuing Justice*, p. 100-2.

equality. In D3, the talented have less primary goods than they do under D2, but the untalented have more than they do under D2. The three distributions can be summarized as follows:

	D1	D2	D3
Talented	d	a	b
Untalented	d	c	b

Table 7.1: *The three distributions in Cohen's critique*

Here letters represent each individuals' shares of primary goods, and $a > b > c > d$.

For D3 to come about, the talented need to be as productive as they are in D2 despite the absence of inequality-generating incentives. In other words, the talented should be committed to an egalitarian ethos that governs their occupational choices. They should decide in which occupations to work and how much to work in light of this egalitarian ethos so that we can co-realize equality and Pareto-efficiency. The Pareto argument, therefore, shows not that we should depart from equality, but that the talented ought to be committed to an egalitarian ethos.

In this chapter, I examine whether an egalitarian ethos is a good idea when the metric of justice is primary goods. Cohen offers an egalitarian ethos as a means of avoiding conflicts between Pareto-optimality and equality, so the desirability of an egalitarian ethos rests on the desirability of the co-realization of Pareto-optimality and equality. In section 7.2, I argue that

Pareto-optimality in the space of primary goods, when the relevant primary goods are income and wealth, is not morally desirable. If the space in which we should seek, or allow, Pareto-optimality and the space in which we seek equality need to be the same, then the argument of section 7.2 is a strong argument against primary goods, and in favour of welfare, as the metric of justice. I examine the case for an egalitarian ethos when the space of equality and Pareto-optimality are welfare in the next chapter. So, in the remainder of this chapter, I explore the possibility of combining the commitment to primary goods as the metric of justice with commitment to Pareto-optimality in another space. I argue that while a commitment to welfare as the space of Pareto-optimality is incompatible with Rawls's commitment to neutrality, the Pareto principle which requires *allowing* Pareto optimality in the space of actual preference satisfaction may not be. The adoption of this Pareto principle, which requires allowing for Pareto improvements in the space of actual preference satisfaction, has significant consequences in Cohen's standard case. In Cohen's standard case, which he identifies as the focus of his critique, the talented prefer D1 over D3, and prefer D2 over D1.[7] On this assumption, the move from D1 to D2 is a Pareto improvement, and mandated by the Pareto principle, whereas the move from D1 to D3 is not a Pareto improvement, and therefore is not mandated by the Pareto principle.[8] Since the Pareto principle doesn't require bringing about D3, an egalitarian ethos aimed to bring about D3 isn't required by the Pareto principle.

[7]Cohen, *Rescuing Justice*, p. 102-3.

[8]This statement needs to be qualified in order to take the preferences of the untalented into account. I discuss this point more fully in section 7.5.

When we reject the Pareto principle that requires Pareto optimality in the space of primary goods, and reflect on the fact that different people prefer equal distributions of primary goods at different absolute levels, the question of the absolute level at which the initial equality should be pitched arises. In section 7.3, I offer a procedure for determining where to set this initial equality and which moves from equality are justified. I then illustrate this proposal with a series of examples (section 7.4). In section 7.5 I examine how this procedure relates to the argument from the original position and the difference principle as it is usually interpreted and point to one unresolved difficulty. In section 7.6, I examine how the Pareto argument for inequality, as it's reconstructed here, handles Cohen's interpersonal test. Finally, in section 7.7, I examine whether Cohen's discussion of the use of inconsistent metrics in the Pareto argument for inequality provides a response to the argument of this chapter.

A few caveats. Firstly, I do not seek to defend the basic structure objection, which holds that occupational choices fall outside the basic structure of society and Rawls's principles of justice applies only to the basic structure. My reconstruction of the Pareto argument for inequality does not rely on this objection. Secondly, my argument applies only to what Cohen has called the standard case. The standard case, as we noted, is the case Cohen identifies as the focus of his critique. Finally, my argument is not intended to be exegetical. I don't claim that I'm elaborating or unearthing an argument that Rawls has offered.

7.2 Metrics of Justice and Pareto Optimality

Here's a common way to define the Pareto principle and some of the related terms

> *Welfare or well-being relative Pareto Optimality (W-Pareto Optimality)* State A is strongly Pareto-superior to state B if everyone is better off in A than in B, and weakly Pareto-superior if at least one person is better off and no one is worse off. State A is Pareto-optimal if no state is Pareto-superior to A. The Pareto principle mandates a Pareto-improvement whenever one is feasible.[9]

In this formulation better off is intended to mean better off in terms of welfare.

Here's another equally common way to define the Pareto principle and some of the related terms.

> *Preference relative Pareto Optimality (P-Pareto Optimality)* State A is strongly Pareto-superior to state B if everyone prefers state A to state B, and weakly Pareto-superior if at least one person prefers state A to state B and nobody prefers state B to state A. State A is Pareto-optimal if no state is Pareto-superior to A. The Pareto principle mandates a Pareto-improvement whenever one is feasible.

[9]Note that I'm following Cohen's formulation very closely in all three versions of the Pareto principle. See Cohen, *Rescuing Justice*, p. 87-8, n. 4.

These two definitions are often used interchangeably by economists, because they identify well being with preference satisfaction.[10] However, these two principles are different in important ways.[11] The W-Pareto principle and the P-Pareto principle are both appealing principles, but with different justifications and orientations. The W-Pareto principle expresses a minimal commitment to promoting human welfare. It says that if we can increase human welfare without harming anyone than we should do so. The P-Pareto principle expresses a commitment to letting people decide their own fate. Even if we think the correct account of welfare is not preference satisfaction, the P-Pareto principle remains intuitively appealing. The P-Pareto principle flows from respect of persons as agents. The basic idea is that even if people's choices will not promote their welfare they should be respected as authors of their decisions. The P-Pareto principle says that the decisions of a group should reflect the wills of its members and if a change is not opposed by anyone, and favoured by some, then that change should be allowed. We can think of the P-Pareto principle as anti-paternalism applied to group decision-making.

Whereas W-Pareto optimality describes a desirable state of affairs to be aimed at, the P-Pareto principle identifies a condition of respecting others as agents. It is about what people should be *allowed* to bring about. So,

[10] Hausman and McPherson also make this observation. See Daniel M. Hausman and Michael S. McPherson, *Economic Analysis, Moral Philosophy, and Public Policy*, 2nd edition. (New York: Cambridge University Press, 2006), p. 65.

[11] Broome makes a similar distinction between what he calls the principle of personal good and the democratic principle. Broome's first corresponds to the W-Pareto principle, and the second corresponds to the P-Pareto principle. See John Broome, *Weighing Goods: Equality, Uncertainty and Time*, (Oxford: Basil Blackwell, 1991), p. 155-159.

even if the correct account of welfare is actual preference satisfaction, the P-Pareto principle and the W-Pareto principle would remain distinct due to their different justifications.

We can introduce a third formulation of the Pareto principle and its related terms, this time relative to the metric of justice.

> *Metric of justice relative Pareto optimality (J-Pareto Optimality)* State A is strongly Pareto-superior to state B if everyone has more of the metric of justice (primary goods, welfare, or capabilities, etc.) in state A than in state B, and weakly Pareto-superior if at least one person has more of the metric of justice and nobody has less of the metric of justice in state A than in state B. State A is Pareto-optimal if no state is Pareto-superior to A. The Pareto principle mandates a Pareto-improvement whenever one is feasible.

Here's an example of a J-Pareto improvement. Suppose that in Distribution A everyone has 10 units of primary goods, and the metric of justice is primary goods. In Distribution B one person has 11 units of primary goods and everyone else still has 10 units of primary goods. Moving from Distribution A to Distribution B would be a weak J-Pareto improvement.

We need two more definitions to explore the relationship between these three formulations of the Pareto principle. We can identify two rival positions on what the proper metric for making comparisons between persons for the purpose of justice is. According to *welfarism*:

> '... in a theory of distributive justice the relevant measure of the opportunities and resources and liberties made available to an individual is the welfare or well-being or utility... that accrues to her from this allotment of goods or that the allotment enables her to achieve.'[12]

Resourcism is the denial of welfarism.[13]

Rawls's theory of justice is an example of resourcism where primary goods are the metric of justice. Primary goods are defined as those goods which everyone is presumed to want, or as those goods which everyone prefers having more of rather than less.[14] Rawls's principles of justice govern the distribution of social primary goods: rights, liberties, and opportunities, and income and wealth, and the social-bases of self respect.[15] The difference principle governs the distribution of income and wealth. In later works, Rawls has suggested that leisure may be included among primary goods, but has not committed himself on this point.[16] Throughout this chapter, I shall assume that leisure is not among the primary goods.

In the case of welfarist theories of distributive justice, J-Pareto optimality requires W-Pareto optimality. Accordingly, the J-Pareto principle can

[12]Richard Arneson, 'Welfare Should Be the Currency of Justice', *Canadian Journal of Philosophy*, 30 (2000):4, p. 502.

[13]Arneson, 'Welfare Should Be the Currency of Justice', p. 502.

[14]John Rawls, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999), p. 79.

[15]Rawls, *Theory*, p. 54; John Rawls, *Political Liberalism*, New edition. (New York: Columbia University Press, 1996), p. 181.

[16]Rawls, *Political Liberalism*, p. 181; John Rawls, *Justice as Fairness: A Restatement*, (Cambridge, Mass: Harvard University Press, 2001), p. 179.

be justified by a commitment to promoting human welfare. However, when the metric of justice is resourcist, the same rationale does not exist. In fact, a J-Pareto improvement may entail a loss of welfare and come into conflict with the W-Pareto principle. This can be illustrated by reference to primary goods. (Since I'm interested in the role of the Pareto principle in the argument for the difference principle, I'm only concerned with the primary goods that fall under the purview of the difference principle, namely income and wealth. Therefore, unless otherwise indicated, when I talk of J-Pareto optimality in the rest of this chapter, what I'm referring to is Pareto optimality in the space of income and wealth. Similarly, unless otherwise indicated, when I refer to primary goods, I'm referring to income and wealth.) There are goods other than primary goods and having more primary goods may require having less of these goods, which in turn may bring about a loss of welfare. Let us imagine a society that is J-Pareto optimal. For this society to be J-Pareto optimal, everyone would have to be as productive as possible. They'd have to work the maximum number of hours they can work, and work in the occupations in which they make the most money. If these two conditions aren't fulfilled, there would be someone who could be more productive and make others, or herself, better off in terms of income and wealth without thereby making anyone else worse off in the same terms. When the society is J-Pareto optimal, its members would have too little leisure. Given reasonable empirical assumptions, this would bring down the welfare of members of that society on any reasonable conception of welfare. Accordingly, a J-Pareto

optimal distribution would be a W-Pareto suboptimal distribution.[17]

The P-Pareto principle can also conflict with the J-Pareto principle. This can again be illustrated using leisure and primary goods. Assume that we are at the J-Pareto optimal distribution we looked at in the previous paragraph. Presumably, at this distribution people would prefer having more leisure. Accordingly, the P-Pareto principle would require departing from this distribution, because there's an outcome which affords more leisure that is a P-Pareto improvement. If we block this move to a P-Pareto superior outcome, we would be overriding the will of individual members of that society.

One objection to the above argument showing the possibility of conflict between P-Pareto optimality and J-Pareto-optimality could be based on Rawls's definition of primary goods as 'things that every rational man is presumed to want'.[18] This, however, would be too quick. It does not follow from this definition that there aren't situations where rational individuals would prefer to have less of primary goods than more of them when trade-offs need to be made between primary goods and other goods. Of course,

[17]Here's how Howe and Roemer, who also make a similar observation, describe the J-Pareto optimal outcome: 'When society assigns the maximin tax scheme, calculated from the unrestricted class of feasible schemes, everyone is miserable. In the case of nonpliable preferences (for instance), each agent is indifferent between receiving his assigned bundle of income and leisure, and starvation.' Roger E. Howe and John E. Roemer, 'Rawlsian Justice as the Core of a Game', *The American Economic Review*, 71 December (1981):5, p. 892. On Howe and Roemer's account someone has pliable preferences if she will prefer any bundle of income and labour to no income and no work. Howe and Roemer summarize the condition of pliable agents as follows: 'Speaking colorfully, one might say the agents will do anything to avoid starvation, whence the adjective "pliable".' Howe and Roemer, 'Rawlsian Justice', p. 886.

[18]Rawls, *Theory*, p. 54.

primary goods can be converted into other goods, but not to all of them, and not by everyone at the same time. Therefore, we need to interpret the standard formulations of primary goods with a *ceteris paribus* clause: People always prefer having more primary goods to having less all else being held equal.

Another possible response to my argument that J-Pareto optimality can conflict with W-Pareto optimality and P-Pareto optimality would be that the problem stems from an unjustifiably narrow understanding of primary goods. It may be thought that if we include leisure and other candidates, which could create such a conflict in our list of primary goods, the problem should not exist. This solution does not work, because the problem would reappear at the level of aggregating these different goods into a single measure for the purpose of interpersonal comparisons. We'll need to give weights to the different goods which make up primary goods in order to compare one person's share with another person's. These weights may differ from the weights these goods have in an individual's utility function.[19] Accordingly a J-Pareto optimal outcome need not be P-Pareto or W-Pareto optimal even on this extended account of primary goods.

The following simple and highly artificial example can be used to illustrate this possibility. Let's say there are only two goods: leisure and income, and we include both goods among primary goods. In order to find out how different individuals compare we would need to devise an index which gives

[19]This point applies to both subjective and objective welfare functions of individuals.

weights to these two goods. Let's say the function we use is the following: $p(w, l) = w + l$ where w is a person's weekly wages and l is the number of hours a person has for leisure, that is the number of hours in the week minus the number of hours of labor. For instance, a person who works 40 hours and receives £200 would end up with 328 units of primary goods. If he were to work 60 hours at the same wage rate, he'd receive £300, and end up with 408 units of primary goods. Assuming that everyone else's condition remains the same a move from the first outcome to the second would be a J-Pareto improvement where the metric of justice is this expanded list of primary goods. Now, suppose this individual's utility function is the following: $u(w, l) = w + 7l$, where w and l are defined as before. In this case, his utility in the first outcome would be 1096. In the second outcome his utility would be 1056. Assuming that everyone's condition remains the same in both outcomes, the second outcome would be P-Pareto inferior even though it is J-Pareto superior.

I have suggested that J-Pareto optimality cannot be justified by an appeal to the P-Pareto principle or the W-Pareto principle when the metric of justice is primary goods. I have also suggested that J-Pareto optimality is likely to come into conflict with both P-Pareto optimality and W-Pareto optimality. Given the significant intuitive appeal of the P-Pareto principle and the W-Pareto principle, this establishes a strong case against the J-Pareto principle when the metric of justice is primary goods. Once it has been granted that, after a certain point, the steps required for generating more income and wealth would result in a loss of welfare, and these steps would be opposed

by individuals, it is difficult to see how a case can be made for J-Pareto optimality. What else other than people's preferences or welfare can motivate the concern with income and wealth? So I suggest rejecting the J-Pareto principle when the relevant goods are income and wealth.

If the J-Pareto principle should be rejected, can the Rawlsian adopt the P-Pareto principle or the W-Principle ? The W-Pareto principle is incompatible with Rawls's commitment to neutrality. We would have to assume that there is a conception of human welfare that the state should seek to promote. Rawls holds that even holding human welfare to be actual preference satisfaction is incompatible with neutrality.[20] He writes:

> '... [T]he state can no more act to maximize the fulfilment of citizens' rational preferences, or wants (as in utilitarianism), or to advance human excellence, or the values of perfection (as in perfectionism), than it can act to advance Catholicism or Protestantism, or any other religion. None of these views of the meaning, value, and purpose of human life, as specified by the corresponding comprehensive religious or philosophical conceptions of the good, are affirmed by citizens generally, and so the pursuit of any one of them through basic institutions gives the state a sectarian character.'[21]

[20] For an elaboration of this idea see Charles E. Larmore, *Patterns of Moral Complexity*, (Cambridge: Cambridge University Press, 1987), p. 48.

[21] John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 453. See also Rawls's discussion of interpersonal comparisons of utility in sections 5 and 6 of 'Social Unity and Primary Goods'.

Even though the W-Pareto principle conflicts with Rawls's commitment to neutrality, even when welfare is preference satisfaction, the P-Pareto principle does not conflict with Rawls's commitment to neutrality. As you will recall the P-Pareto principle is justified by appeal to respecting the wills of individuals rather than promoting their welfare. Suppose we have two distributions Distribution A and Distribution B. Suppose there is at least one person who prefers Distribution A to Distribution B, and no one prefers Distribution B over Distribution A. Under these circumstances, a social decision procedure which respects the wishes of those involved should bring about, or not stay in the way of, Distribution A. This is, however, not because Distribution A is thought to bring about an increase in welfare. By saying that Distribution A should come about we're not making assumptions about what the true account of well-being is. We are just respecting the choices of individuals.

Given the problems faced by W-Pareto optimality and J-Pareto optimality in the Rawlsian framework, and the intuitive appeal of the P-Pareto principle, the Rawlsian should employ the P-Pareto principle in the second stage of the Pareto argument for inequality. On this proposal, we will be evaluating individuals' shares in one space (primary goods) and pursuing, or, more accurately, allowing, Pareto-optimality in another space (actual preferences). The pursuit of equality in one space and of Pareto-optimality in another may seem problematic. I have argued that there are good reasons for not pursuing Pareto optimality in the space of primary goods. So, if it is shown that there is something wrong with the pursuit of equality in one space and of Pareto-

optimality in another, this, coupled with the argument against the pursuit of Pareto optimality in the space of primary goods, would be an argument against using primary goods as the metric of justice, and an argument in favour of welfare as the metric of justice.

Since we shall look at the case for egalitarian ethi when welfare is assumed to be the metric of justice in the next chapter, and it's not obvious that the metric of justice and the space in which we should pursue Pareto optimality ought to be the same, I'd like to explore this proposal, which holds that the metric of justice is primary goods but we should be interested in P-Pareto optimality rather than J-Pareto optimality. Let me also offer a few reasons for why this proposal is not obviously incoherent or out of step with the rest of Rawls's theory. Firstly, the argument that J-Pareto optimality can conflict with P-Pareto optimality and W-Pareto optimality does not conflict with Rawls's account of primary goods. Rawls is clear that primary goods are not intended as a measure of a person's well-being.[22] Rawls also implicitly concedes that J-Pareto optimality can conflict with P-Pareto optimality when he writes:

> 'It is a mistake to believe that a just and good society must wait upon a high material standard of life. What men want is meaningful work in free association with others... within a framework of just basic institutions... To achieve this state of things great wealth is not necessary. In fact, beyond some point it is more likely to be a positive hindrance, a meaningless distraction at

[22]Rawls, *Collected Papers*, pp. 370, 455.

best if not a temptation to indulgence and emptiness.'[23]

It's worth emphasizing that the primary goods for which we aren't pursuing J-Pareto optimality are only income and wealth. Considerations of P-Pareto optimality are not allowed to override equal distributions of basic liberties, and basic liberties are lexically prior. Furthermore, the liberty principle guarantees the *fair value* of liberties. This ensures that many of the possibly objectionable outcomes of bringing preferences into the picture are blocked.

Furthermore, in cases where we have resourcist intuitions, the P-Pareto principle which requires allowing for P-Pareto improvements seem intuitively justified. For instance, the following set of attitudes seem natural for parents adopt.

> If one's children are roughly the same age and of a certain level of maturity they should have the same allowance. (An analogue of equality of resources)
>
> Giving children more allowance is not necessarily a good thing. (An analogue of the rejection of J-Pareto optimality)
>
> Provided certain conditions are fulfilled, if the children agree on a deal between them which results in an inequality between them they should be allowed to make that deal (An analogue of the P-Pareto principle)

Finally, a commitment to the P-Pareto principle does not entail giving preferences a role in interpersonal comparisons. When we compare people's

[23]Rawls, *Theory*, p. 257.

shares, we will be looking at their shares of primary goods, and we will not be concerned with their preferences. When the P-Pareto licences an inequality, it doesn't licence the inequality because the person who ends up with more income gets less utility from income than the person who ends up with less income. The P-Pareto licences the inequality, because both individuals prefer the unequal outcome.

Let me digress for a moment to fulfil my promise to address the relationship between Pareto optimality and justice that I made in section 4.6. Are P-Pareto optimality and W-Pareto optimality prima facie irrelevant from the perspective of justice? As you will recall we are only interested in Cohen's claim that, as a conceptual matter, considerations of Pareto optimality are alien to justice. The P-Pareto principle is about people being allowed to bring about outcomes that are not opposed by anyone. Within certain boundaries, which will be matters of contention, people seem to have a right to bring about such changes. A theory of justice will set out these boundaries, but it will allow for room for this right. The P-Pareto principle is a natural part of a theory of justice, which lets some matters be decided by actual persons, and for that reason is not prima facie irrelevant to justice. Of course, it can be claimed that a theory of justice should not leave any matters to be decided by the choices of actual persons, but that would be a substantive issue. Those who believe that some matters are to be decided by the choices of actual persons are not displaying a misunderstanding of the concept of justice. The W-Pareto principle, we said, expresses a commitment to promoting human welfare. It says that if person's welfare can be increased without thereby

decreasing anyone else's welfare we should do so. We could certainly claim that laws that sought to equally minimize the welfare of individuals were unjust, because they failed to show due concern for individuals. To claim that people can expect as a matter of right that their laws should seek to maximally promote the welfare of each individual, and when it is not possible to do this consistently with assuring everyone of equal amounts of welfare, we should allow for improvements that do not harm anyone, does not betray a misunderstanding of the concept of justice.

7.3 Looking at Cohen's Critique Again

With the distinctions between different types of Pareto-optimality and our argument that Rawls should favour P-Pareto optimality, let's look at Cohen's criticism of the Pareto argument for inequality again.

Cohen does not distinguish between P-Pareto optimality and J-Pareto optimality. However, from his discussions we can see that he has J-Pareto optimality in mind. When introducing D3, Cohen writes 'D3 is Pareto-superior to D1'.[24] And in summarizing his argument Cohen writes:

> '... [W]hen D2 is possible, then so, standardly, is D3, a Pareto-optimal state of equality which is Pareto-superior to D1 and Pareto-incomparable with D2. We therefore do not have to choose between equality and Pareto.'[25]

[24] Cohen, *Rescuing Justice*, p. 183
[25] Cohen, *Rescuing Justice*, p. 183

However Cohen wants the main focus of discussion to be the standard case. In Cohen's standard case the talented's preferences are as follows, D2>D1>D3. In other words, the talented prefer D1 over D3. Under these circumstances, D3 is not P-Pareto-superior to D1, but it is J-Pareto superior. Accordingly, we can conclude that Cohen has J-Pareto optimality in mind.

We have seen that the Rawlsian should drop the J-Pareto principle, and adopt the P-Pareto principle. But, in Cohen's standard case, the P-Pareto principle is indifferent between D1 and D3. Assuming that the untalented prefers D3 over D1, and the talented prefers D1 over D3, the two distributions are Pareto-incomparable. Since both D1 and D3 are equal distributions equality is also indifferent between D1 and D3. Therefore, the superiority of D3 has not been established. Even though the P-Pareto principle is indifferent between D1 and D3, it is not indifferent between D1 and D2. We already know that the talented prefer D2 over D1. If we also assume that the untalented also prefer D2 over D1, then a move from D1 to D2 is a strong P-Pareto improvement and recommended by the P-Pareto principle.[26] Cohen's proposed solution that preserves Pareto optimality and equality only works when we assume that we are interested in J-Pareto optimality. However, if we're interested in allowing P-Pareto optimality, then we still have to choose between equality and Pareto optimality.

This suggests a new interpretation of the Pareto argument for inequality, which I shall outline for a two person case. I shall outline two versions

[26]There are some valid worries about whether the untalented will also prefer D2 over D1. I shall come back to them in section 7.5.

of this interpretation. The first version is more general and justifies the difference principle along with several other principles. The second version is less general. It relies on two additional premises, but picks the difference principle uniquely.

Consider a series of equal distributions of primary goods starting from d_0 to d_n, where each successive distribution contains more primary goods than the one that comes before it. d_0 is the distribution where no one has any primary goods and d_n is the distribution where everyone has the maximum amount of primary goods that is feasible. We assume that d_n includes those distributions which would be feasible if we had complete knowledge of each individual's skills and utility functions and we could impose lump sum taxes based on these. We also assume that d_n includes those distributions which could only be brought about by an egalitarian ethos. The first stage of the Pareto argument for inequality is assumed to establish a presumption in favour of an equal distribution of primary goods. Since all of these distributions are equal, the first stage of the Pareto argument for inequality does not give us reason to prefer any one of these distributions. We have seen that J-Pareto optimality is not desirable. Accordingly, we also don't have a reason to pick d_n.

The P-Pareto principle, as distinct from W-Pareto principle, also does not single out one of these distributions as the one we should aim for despite the fact that one of these equal distributions may be P-Pareto efficient. The P-Pareto principle does not describe a valuable state of affairs to be aimed at.

It spells out a condition which is necessary for respecting others as agents. From the perspective of the P-Pareto principle, what matters is not whether a distribution is P-Pareto efficient or not, but whether social institutions allow for P-Pareto improvements. Accordingly, even if one of these distributions is P-Pareto efficient, the P-Pareto principle does not give us a reason to favour it.

An analogy with the right to freedom of expression may be instructive at this point. A social system may provide everyone with an effective right to freedom of expression. However, people may choose not to exercise this right. What matters from the perspective of the right to freedom of expression is that people are allowed to express their opinions. Similarly, the P-Pareto principle says that people should be allowed to bring about certain outcomes, not that P-Pareto improvements have to be brought about. The analogy is strained by the fact that people act on their preferences, and therefore what the changes that P-Pareto principle says ought to be allowed will usually come about. However, there can be cases where people aren't aware of possible P-Pareto improvements. Such cases aren't causes of concern from the perspective of the P-Pareto principle.

Everyone is assumed to be entitled to an equal share of primary goods (the first stage of the Pareto argument for inequality), and they are also assumed to be allowed to bring about certain outcomes they prefer provided no one has any objections (the P-Pareto principle). Since different equal distributions of primary goods contain different bundles of income and leisure,

and people have different preferences over these goods, they prefer different equal distributions. Despite disagreeing about the desirability of some equal distributions at different absolute levels, they may agree that some equal distributions are preferable to others. This suggests the following procedure for arriving at a favoured distribution.

> *The n-step procedure* We start from d_0 and ask both individuals whether they prefer d_1 to d_0. If they both prefer d_1 to d_0, then we move to d_1. We repeat this procedure for each d_i and d_{i+1} until there is a person who does not prefer d_{i+1} to d_i. We identify this distribution as D1. Both individuals are guaranteed as much primary goods as they receive under D1. According to the P-Pareto principle, if there is an unequal distribution which both individuals prefer to D1 and agree on, we should allow these individuals to move to that unequal distribution, which we identify as D2.

There are five problems with the n-step procedure. The first problem is that we can just as well start an analogous procedure from d_n. We could ask everyone whether they prefer d_{n-1} to d_n and move to d_{n-1} if they do. We could repeat this process until we identify the point where a move from d_i to d_{i-1} is not wanted by everyone. We could set this as the equal distribution that both individuals are guaranteed. This is a problem, because the unequal outcome the parties would agree on is not independent of the equal distribution they have been guaranteed. As a result, the procedure would pick a different unequal outcome depending on whether we begin the procedure from d_0 or d_n. The second problem arises in the following way. For most

specifications of equal distributions of primary goods, there corresponds more than one distribution of goods which are not primary goods such as leisure and job satisfaction. Which of these is realized can also influence whether the parties agree to a move or not. The third problem is that this procedure would be very difficult to carry out when the number of people increases. The fourth problem is that there will be several outcomes which would be P-Pareto improvements over the outcome identified in the first stage of this procedure, which the parties could agree on. The procedure does not identify a unique P-Pareto efficient distribution. The fifth problem is that the veto power individuals enjoy is too wide-ranging. Individuals have a veto power not only over the use of their own productive capacities, but also those of others. Even if one individual would like to work at his full capacity, the other can stop him from doing so.

Due in particular to the second reason mentioned above, the n-step procedure is unsuitable for cases where there is production. Nevertheless there are cases where it can have application. Here is one such case. Suppose there are two street fair organizing committees in a city (Committee A and Committee B). There are five streets on which fairs can be organized, and it is not possible to have two different fairs on the same street. The council will allocate the streets to both committees, and the committees are required to run fairs on the streets they have been allocated. We begin by asking the heads of both committees whether they would like to be allocated a street each. They answer yes, so we ask them whether they would like to be allocated two streets each. We learn that even though Committee B is in favour

of this move, which allocates both committees two streets each, Committee A opposes it. Committee A thinks that having fairs on four streets would be too disruptive. Since Committee A has vetoed the move to the distribution that assigns both committees two streets each, the distribution that assigns both committees a street each is set as the initial equal distribution. Suppose that Committee A thinks that having fairs on only three streets would not be too disruptive. In this case, there's a possible P-Pareto improvement: granting only one of the committees an extra street. The distribution that both committees agree to will be the final distribution identified by the n-step procedure.

We need a non-arbitrary way of determining which equal distribution we identify as the initial distribution guaranteed to both individuals, a way of picking a unique P-Pareto efficient outcome, and limiting the range of the veto power of the individuals involved. As a possible solution to these difficulties consider the following proposal.

> *The two step-procedure* Suppose we have two agents. They know that primary goods will be distributed equally. They know each other's utility functions and each other's productive capacities. In light of this information, they make their occupational choices. The resulting equal distribution of primary goods is where we set the baseline equality. Assume that this distribution is P-Pareto inefficient. The n-step procedure allowed these two individuals to bargain over the alternatives and pick the distribution they both agreed to. Instead, we now pick, among the P-Pareto im-

provements available, the distribution which makes the worst off as well off as possible.

There is room for strategic manipulation in the two-step procedure which needs to be explicitly ruled out. When making their first occupational decision, both individuals should make their decision as if the equal distribution was final. Their choice should reflect their true preferences. Otherwise, by misrepresenting their preferences in the first stage, individuals can ensure that they end up in an advantageous position in the second stage of the procedure.

The first change the two-step procedure introduces is in how the non-agreement point is identified. Instead of going through each successive equal distribution, we pick the equal distribution which comes about through the occupational decisions of individuals. This way of determining the non-agreement point provides a solution to the first three problems I identified with the previous procedure. The decisions of individuals is consistent with their commitment to the first stage of Barry's argument since they are merely acting in ways which will realize their favoured equal distribution of primary goods. Furthermore, this procedure provides one appealing answer to the question: 'If equality of primary goods is required and different people prefer different equal distributions, which one should we pick?' We should pick the equal distribution that would be brought about by the joint choices of everyone involved. The two-step procedure also limits the veto power of individuals. Their veto power now extends only over their own productive

capacities. Picking the P-Pareto improvement which makes the worst-off as well-off possible provides a solution to the fourth problem identified above, that of identifying a unique distribution among the possible P-Pareto improvements. This way we are picking the P-Pareto improvement which is closest to equality.

7.4 The Two-Step Procedure Illustrated

Let us look at a few cases to illustrate how the two-step procedure operates, and show some of its implications which may not be immediately clear.

Imagine a very simple scenario with only two people, Mark and Jane. Initially, there is only one type of work Mark can do (Job L). Job L produces 15 widgets per day. We shall assume that widgets are primary goods. We shall also assume that there are no other primary goods under consideration. Jane is more skilled. She can choose between two different jobs. She can either do Job L, or she can do Job P which produces 27 widgets per day. For simplicity we assume that the only occupational choice Mark and Jane face is which job to work in. We do not consider the possibility of Mark and Jane altering the number of hours they work.[27] We assume that each widget produces one unit of utility for both Mark and Jane, and counts as one unit of primary goods. We do not assume that utility is interpersonally comparable.[28]

[27]The choice in working in a more or less productive occupation could alternatively be interpreted as the choice between working more or less hours. Such reinterpretation does not alter the cases.

[28]I ignore diminishing marginal utility for the sake of simplicity.

Mark and Jane do not end up with what they produce. There is an egalitarian distributor, who divides up the widgets they produce so that they end up with equal shares of primary goods. We assume that Mark and Jane know each other's utility functions and productive capacities. We also assume that they make their choices with the aim of maximizing their utility. D3 refers to the distributions in which Mark and Jane end up with the maximum amount of primary goods that is feasible. D1 refers to the distribution which will come about when Jane and Mark make occupational decisions in light of their knowledge that primary goods will be distributed equally.

Our scenario is in two respects relevantly similar to the situation in a Rawlsian society. Firstly, Mark and Jane are not guided by an ethos in their occupational choices. They make their choices solely with their own interests in mind. Secondly, they comply with the directives of the social institutions which govern the distribution of primary goods.

Case A. Let us initially assume that widgets are the only source of utility in Mark and Jane's lives, and work has no disutility. (I shall later drop some of these assumptions.) Supposing that Mark and Jane are both self-interested and seek to maximize their utility, which job should they choose? We are taking for granted that each agent has decided to work, and we know Mark can only do Job L. For this reason he will produce 15 widgets. Jane can produce either 15 or 27 widgets. If she produces 15 widgets, they will both end up with 15 units of primary goods and 15 units of utility. If she produces 27 widgets, they will both end up with 21 units of primary goods

and 21 units of utility. Mark and Jane's payoffs are given below.[29]

Table 7.2: *Case A*

	Job L	Job P
Job L	15P/15U, 15P/15U	21P/21U, 21P/21U

Given that her utility is maximized when she does Job P, Jane will choose to take up Job P, thereby making Mark also better off. Under these highly restrictive conditions, the more skilled person, even though solely motivated by self-interest in her choice of occupation, will act in a way that will maximize everyone's share of primary goods even though primary goods are going to be distributed equally. Even though the more skilled individual is not motivated by an egalitarian ethos in her occupational choice, the resulting distribution is equal and both J-Pareto efficient and P-Pareto efficient.

Case B. Now suppose that Jane has an aversion to Job P. Job P has 9 units of disutility for Jane. Under equality of primary goods, if Jane chooses Job P, she will end up with 21 units of primary goods, and 12 units of utility. Mark will have 21 units of primary goods and 21 units of utility. If Jane chooses job L, she will end up with 15 units of primary goods and 15 units of utility. Mark and Jane's payoffs are given below.

[29]In this and the following, tables, the columns represent Jane's choices. The rows represent Mark's choices. Mark and Jane's payoffs are given both in terms of primary goods (P) and utility (U). The figures before the comma are Mark's payoff. The ones after are Jane's. Utilities are assumed to be cardinal. They are not assumed to be interpersonally comparable.

Table 7.3: *Case B*

	Job L	Job P
Job L	15P/15U, 15P/15U	21P/21U, 21P/12U

In this case, Jane will choose Job L. This will be the initial equal distribution identified by our two-step procedure. There are P-Pareto improvements over this outcome. Outcomes which give both Jane and Mark more than 15 units of utility are also feasible. In other words there are outcomes which both Mark and Jane prefer to this first outcome. Among these P-Pareto improvements, our procedure picks the one which gives Jane slightly more than 24 units of primary goods and gives Mark slightly less than 18 units of primary goods. The procedure picks the distribution which gives Jane slightly more than 24 units of primary goods, because, among the P-Pareto improvements over D1, that is the distribution which makes Mark as well off as possible in terms of primary goods.

We can contrast this outcome with what we can call *laissez-faire* where both individuals get to keep what they have produced. Under laissez-faire, Jane would choose Job P, and end up with 27 units of primary goods and 18 units of utility; Mark would end up with 15 units of primary goods and 15 units of utility.

Case C. Suppose Jane finds Job L to be particularly enjoyable. Just doing Job L provides her with 15 units of utility. In this case, Mark and Jane's

payoffs are as follows.

Table 7.4: *Case C*

	Job L	Job P
Job L	15P/15U, 15P/30U	21P/21U, 21P/21U

Job L maximizes Jane's utility. Accordingly, Jane will choose Job L. In this case, there are no P-Pareto improvements over D1, and the second step of the two-step procedure is not applicable. In this instance, Mark and Jane's shares are the same as what they would end up with under laissez faire.

Suppose Jane was committed to bringing about J-Pareto optimality, and chose Job P. This outcome would not be P-Pareto efficient. There are outcomes preferred by both Mark and Jane to D3–D3 being the outcome in the right hand column.

The possibility of P-Pareto improvements over D3 gives us an opportunity to illustrate the first problem that I raised in my discussion of the n-step procedure. If we set D3 as the non-agreement point and moved to a P-Pareto efficient outcome from there, then Mark and Jane's shares would be different. The P-Pareto improvements over D3 are the outcomes which give Mark more than 21 and less than 24 units of primary goods, and give Jane more than 6 and less than 9 units of primary goods. In this outcome Jane would be working in Job L.

Case D. So far, we assumed that only Jane had a job choice. Now, let us suppose that Mark also has a choice. Mark faces a choice between Job L and a job which is even less productive, producing only 3 widgets per day (Job E). Mark enjoys job E immensely and just doing it is worth 9 units of utility for him. Neither job has any utility or disutility for Jane. Mark and Jane's payoffs are given below.

Table 7.5: *Case D*

	Job L	Job P
Job L	15P/15U, 15P/15U	21P/21U, 21P/21U
Job E	9P/18U, 9P/9U	15P/24U, 15P/15U

Jane will choose Job P which maximizes her utility. Mark will choose Job E. As a result, the outcome in the lower right hand corner will be realized.

D1 is not P-Pareto optimal. There are outcomes in which both Mark and Jane do better in utility terms than they do in D1. These outcomes are the ones in which Mark gets more than 24 and less than 27 units of primary goods, and Jane gets more than 15 and less than 18 units of primary goods. Of these P-Pareto-improvements, our two-step procedure picks the one which gives Mark slightly more than 24 units of primary goods and Jane slightly less than 18 units of primary good, because this is the outcome which makes Jane as well off as possible.

Mark does considerably better under this scheme than he would under laissez-faire. Under laissez-faire he would end up with 15 units of primary goods and 15 units of utility. Jane does worse under this scheme. Under laissez-faire she would end up with 27 units of primary goods and 27 units of utility.

As this example illustrates, on the two-step procedure, the person who ends up better-off in terms of primary goods is not necessarily the one with greater productive capacities. Mark ends up with more primary goods than Jane, even though Jane has greater productive capacities. On the two-step procedure, for a person to end up better-off, she has to (a) be in a position to increase her production; and (b) prefer her bundle of primary goods and non-primary goods in D1 to her bundle in D3.

7.5 The Difference Principle and the Argument from the Original Position

Now that we have seen how the two-step procedure works, a natural question to ask is how this version of the Pareto argument for inequality relates to the argument from the original position, and the difference principle as it is usually interpreted. If we look at the difference principle by itself, then there's a significant divergence between it and the principle we get with our reconstructed Pareto argument for inequality. The difference principle reads:

'Social and economic inequalities are to be arranged so that they

are... to the greatest benefit of the least disadvantaged...'[30]

Here, of course, benefits are primary goods. The principle seems to require J-Pareto optimality.

If we look at the reasoning in the original position, it also seems to suggest that we should pursue J-Pareto optimality. The denizens of the original position do not know their conceptions of the good. They use their expected shares of primary goods, and the assumption that 'they would prefer more primary goods rather than less' to rank different alternatives. This again suggests that principles of justice would require J-Pareto optimality rather than P-Pareto optimality. The denizens of the original position would rank distributions which afforded them greater shares of primary goods higher even though they may find out, once the veil of ignorance is lifted, that this is an outcome that is not P-Pareto efficient. In fact, there does not seem to be any way for taking the preferences of actual persons into account within this framework.

This apparent clash between the Pareto argument for inequality as I have reconstructed it, and the difference principle can be overcome by observing that even though the difference principle requires pursuing J-Pareto optimality, it requires pursuing J-Pareto optimality within a certain framework. The difference principle applies to the basic structure of society. In addition to this, the liberty principle, which protects freedom of occupational choice, is

[30]Rawls, *Theory*, p. 266.

lexically prior to the difference principle.[31] It is within these constraints that we are required to pursue J-Pareto optimality. Within these restrictions, the pursuit of J-Pareto optimality coincides with the application of the two-step procedure which does not require J-Pareto optimality. To see this point intuitively, consider again, Cohen's critique. Cohen holds that the difference principle should apply to the occupational choices of individuals, and when it does we can realize D3, otherwise we'll be stuck with D2. As I argued, the Pareto argument for inequality requires that we realize D2. So the application of the difference principle within these constraints brings about the result required by our two step procedure.[32]

Let us look again at Case B to elaborate this point. In Case B, Mark and Jane's pay offs were as follows:

Table 7.6: *Case B*

	Job L	Job P
Job L	15P/15U, 15P/15U	21P/21U, 21P/12U

We said that our two-step procedure would recommend the outcome that would be realized when Jane worked in Job P, Mark worked in Job L, Jane

[31] It is not wholly clear whether Rawls wants to include freedom of occupational choice in the liberties ensured by his first principle or whether it should be protected by the fair equality of opportunity. Either way, it's lexically prior to the difference principle. For further discussion see Cohen, *Rescuing Justice*, p. 196-7.

[32] This interpretation of the Pareto argument for inequality also furnishes us with a justification for why the application of the difference principle should be restricted to the basic structure. It is only when there is a constraint on the scope of the difference principle that it brings about morally desirable outcomes.

received slightly more than 24 units of primary goods and Mark received slightly less than 18 units of primary goods.

Under the two constraints mentioned, the main tool available to put the difference principle into practice is an income tax. The tax rate is chosen with the aim of ensuring that the worst off's share of primary goods is maximized. Given this goal, the best case scenario would be to have Jane work at Job P and distribute primary goods equally, that is, to achieve J-Pareto optimality and equality. At the tax rate which would bring about equality, Jane will not choose to work in Job P. What needs to be done is to pick an income tax which will induce Jane to choose Job P, and maximally benefit Mark. In order for Jane to choose Job P, she should end up better off, in utility terms, than she would when she chose Job L. This means that the tax rate we pick should leave Jane with at least 24 units of primary goods when she does Job P, otherwise she will not pick Job P. If we are to maximize Mark's share of primary goods, then Jane should get only slightly more than 24 units of primary goods. We will then be able to give Mark slightly less than 18 units of primary goods. This was the outcome required by our two-step procedure. The difference principle as it is usually interpreted, and under the constraints previously mentioned, mimics the two-step procedure.

The interpretation of the difference principle I've offered departs from Rawls in two important respects. Firstly, we've assumed that the difference principle applies to individuals rather than groups. Secondly, in the income tax we have used in our example, the tax rates are tailored to individuals. If

we do not make this assumption, then we'll end up with different outcomes.

To see how such divergences can occur consider the following case where there is an addition to our cast of characters named Seyla.[33] Suppose facts of this case pertaining to Mark and Jane are the same as Case B. Like Jane, Seyla can work in both jobs, but unlike Jane she's indifferent between the two jobs. The payoff table is as follows.[34]

Table 7.7: *Case E*

	Job L	Job P
Job L	15P/15U, 15P/15U, 15P/15U	19P/19U, 19P/10U, 19P/19U
Job P	19P/19U, 19P/19U, 19P/19U	23P/23U, 23P/14U, 23P/23U

The first distribution identified in the two-step procedure will be the distribution in the lower left-hand corner where Jane works in Job L and Seyla works in Job P. There are Pareto improvements over this outcome. Jane will work in Job P instead of Job L if she gets more utility than 19 units. This can be done by giving her more primary goods than others. To be precise, if she gets slightly more than 28 units of primary goods, she will choose Job P. We would then be able to give Seyla and Mark slightly less than 20.5 units of primary goods. They'd prefer this unequal outcome to

[33]Incidentally, this case also illustrates how our two-step procedure is able to handle cases which have more than two people.

[34]The columns represent Jane's choices. The rows represent Seyla's choices. As before, I've given the payoffs in terms of primary goods (P) and utility (U). The first pay off is Mark's, the second is Jane's and the third Seyla's. Utilities are assumed to be cardinal. We don't need to assume that they are interpersonally comparable.

Jane working in Job L. When Jane works in Job L, they end up with 19 units of utility. In this unequal outcome, they end up with slightly less than 20.5 units of utility. Furthermore, this is the distribution which maximally benefits Seyla and Mark. Accordingly, our two step procedure would require this distribution.

Putting this outcome into practice would require having different tax rates for Seyla and Jane. Jane does not prefer working in Job P unless offered incentives, but Seyla does not need any incentives to work in Job P. Accordingly she could be taxed at a higher rate than Jane and still work in Job P. If Seyla faced the same tax rate as Jane, this would leave Mark worse off than our two-step procedure requires us to.

Case E also shows why my account only applies to Cohen's standard case. In the standard case, the talented will produce less if the tax rate is higher. In the bad case, the talented will not produce less if the tax rate is higher, but they deliberately misrepresent their preferences for strategic reasons.[35] Seyla's preferences are the same as the preferences of the talented individuals in Cohen's bad case. If our two-step procedure required giving incentives to Seyla as well as Jane, it would be sanctioning the bad case as well as the standard case. We have stipulated that when individuals make their choices in the first step of the two-step procedure, they don't act strategically. This ensures that the two-step procedure does not justify giving incentives to the talented in the bad case. When the income tax rate is not tailored

[35]Cohen, *Rescuing Justice*, p. 57-8

to individuals, then both Seyla and Jane end up with a greater share of primary goods than Mark. The two-step procedure requires only Jane to receive incentives, so the tax-rate needs to be tailored to individuals.

Let me mention one unresolved problem with the two-step procedure.[36] In section 7.3, when arguing that D2 would be a strong P-Pareto improvement over D1, I assumed that the untalented also preferred D2 over D1. To support this assumption we can reason as follows. When we move from D1 to D2, the change in the total amount of income and wealth is due to the change in the occupational choices of the talented. So, the untalented gets more income and wealth without any extra labor burden. Therefore a good case can be made that the untalented would also prefer D2 over D1. If the untalented does not care for money, then she would be indifferent between D2 and D1. In which case the move from D1 to D2 would be a weak P-Pareto improvement. Despite, this the untalented may prefer D1 over D2, because in D2 they would envy the better off, or because they believe that D1 contains more solidarity.[37] In this case, the two-step procedure would require staying at D1 whereas the difference principle would require moving to D2. What we need to preserve the fit between the outcome of two-step procedure and what the difference principle recommends is an assumption like the assumption of mutual disinterestedness that Rawls attributes to the denizens of the original position. However, that assumption seems difficult to justify given the two-

[36] I'm grateful to David Miller for pointing out this problem.

[37] Given the lexical priority of the liberty principle, they can't claim that D2 is preferable to D1, because in D2 they would enjoy, for instance, less political influence. Such outcomes are already ruled out, because the liberty principle guarantees the fair value of the political liberties. Rawls, *Political Liberalism*, p. 324-331.

step procedure's commitment to the P-Pareto principle.

7.6 The Interpersonal Test

One thing we want to know is whether the two-step procedure passes Cohen's interpersonal test, which has strong intuitive appeal, particularly for contractualists. The interpersonal test

> '...tests how robust a policy argument is, by subjecting it to variation with respect to who is speaking and/or who is listening when the argument is presented. The test asks whether the argument could serve as a justification of a mooted policy when uttered by any member of society to any other member... If, *because* of who is presenting it, and/or to whom it is presented, the argument cannot serve as a justification of the policy, then whether or not it passes as such under other dialogical conditions, it fails (*tout court*) to provide a comprehensive justification of the policy.'[38]

Cohen gives the example of an argument, which fails the interpersonal test:

> 'Children should be with their parents.
>
> Unless you pay me, I shall not return your child.
>
> So, you should pay me.'[39]

The argument is uttered by the kidnapper who makes the second premise true, and fails the interpersonal test. Cohen, then asks us to imagine the

[38] Cohen, *Rescuing Justice*, p. 42, emphases in the original.
[39] Cohen, *Rescuing Justice*, p. 39.

following argument uttered by managers:

> 'Public policy should make the worst off people (in this case, as it happens, you) better off.
> If top tax goes up to 60%, we shall work less hard, and, as a result, the position of the poor (your position) will be worse.
> So, the top tax on our income should not be raised to 60%.'[40]

Cohen is right to deny that this argument fails the interpersonal test. However, on the interpretation of the Pareto argument for inequality offered here, this is not the argument those who end up better off should make. The argument of someone who will end up better off should be the following:[41]

1. No one has a prior claim to having more primary goods than others.
2. I prefer D1 to D3; you prefer D3 to D1.
3. I can act in ways which will realize D1. This is consistent with my commitment to (1).
4. There may be outcomes that we both prefer to D1, and public policy should allow us to realize those outcomes.
5. Of these outcomes that we both prefer to D1, the tax rates should be set at the point which makes you as well off as possible.

[40] Cohen, *Rescuing Justice*, p. 59.

[41] For ease of exposition, I'm assuming that we're at D1, instead of being already at an unequal distribution.

Those who will end up with more in D2 do not justify their actions by citing the difference principle. The difference principle is an outcome of the argument they employ here (coupled with the argument of section 7.5) rather than a premise in their argument. Furthermore, their acting in ways which will realize D1 is compatible with their belief that they are not entitled to more primary goods than others. They are not striking a posture for bargaining purposes so that they end up with more primary goods than others. They are willing to stick by an equal outcome.

Can't those who end up worse-off in D2 object that those who end up better off in D2 are, already, better off in welfare terms in D1 than they are? Alternatively can't they argue that those who end up better off already have more leisure than others in D1? The line of reasoning behind this could be the following: 'Some people prefer D1 to D3, whereas others prefer D3 to D1. For this to be the case, those who prefer D1 to D3 should be better off in utility terms, or have more of something else, for instance leisure, than those who prefer D3 to D1'.

As you will recall in our examples, we did not need to assume that utilities were interpersonally comparable. No interpersonal results follow from the fact that one person prefers D1 and the other prefers D3. And even if, at D1, those who prefer D3 over D1 were worse off in utility terms than those who prefer D1 over D3, this would not be a legitimate complaint as long as we are assuming that primary goods are the correct metric of justice. If primary goods are the correct metric of justice, then whether someone has

more utility is not a concern for us from the perspective of justice.

The claim in (3) may seem problematic. One possible explanation of seeing (3) as problematic would be the assumption that D3 is somehow privileged. Once we distinguish between J-Pareto Optimality, P-Pareto Optimality and W-Pareto Optimality and reject J-Pareto optimality it should be clear that D3 is not privileged when the metric of justice is primary goods. When we realize that D3 is not privileged, the question of the absolute level at which equality should be pursued when different people prefer equality at different levels arises. The first step of the two-step procedure lets this question be answered by the joint decisions of individuals.

7.7 Cohen on Inconsistent Metrics

Several people, I presented the argument of this chapter to, have suggested that my argument is addressed in the Inconsistent Metrics section in Chapter 2 of Cohen's book. Let me explain why what Cohen says in that section has no bearing on the arguments of this chapter.

In that section, Cohen is objecting to the following line of thought. When we move from D1 to D3, the talented are more productive whereas the work load of others has not changed. Consequently, instead of D3, we should move to an unequal outcome which compensates people for working harder.

As Cohen rightly suggests, this line of reasoning is mistaken because it employs two different metrics. Either labor burden should be included in the

metric of justice, in which case D2 cannot be characterized as an unequal outcome, or labor burden should not be included in the metric of justice, in which case it cannot be used as a reason for preferring D2.[42]

Cohen is quite right to hold that we need to apply the same metric in evaluating individuals' shares in D1, D2, and D3. I have not suggested that those who act in ways which will realize D1 have a legitimate claim to do so because they would be under an uncompensated burden in D3. On the proposal we've been examining, those who end up better off in D2 are justified, because D1 and D3 are symmetrically placed. That is, when the J-Pareto principle is rejected, there's no reason to prefer D3 over D1. I have argued that the first stage of the Pareto argument for inequality does not determine the absolute level at which we should seek equality, and that J-Pareto optimality is not desirable. Accordingly, people who act in ways which will realize their preferred equal distribution of primary goods are not acting in ways that conflict with the first stage of the Pareto argument for inequality.

As I pointed out in the previous section, we can't infer any interpersonal results from the fact that one person prefers D1 over D3 and another prefers D3 over D1. It's possible that the person who prefers D1 over D3 is worse off in welfare terms than the other person in both D1 and D3. We might, nevertheless, be worried that the talented person who prefers D1 over D3 enjoys some unfair advantages.

[42]Cohen, *Rescuing Justice*, p. 108.

Suppose we have two persons, talented, T, and untalented, U. T prefers D1 over D3 and U prefers D3 over D1. Suppose that at D1, T doesn't work at all, and lives off his share of the income generated by U. Let's also assume that D3 is the outcome generated when T works the same number of hours as U does in D1. In such a case, the move from D1 to D2, which gives T more primary goods than U will seem unfair. It is true that this move seems unfair provided there isn't a special explanation of T's aversion to work. But so does D1. The problem this example identifies is with the metric rather than with the move from D1 to D2. The natural response to the unfairness in this case will be to incorporate leisure into the metric. This will, of course, change the correct description of the cases. D1 will no longer count as an equal distribution. However, even on this revised metric, some people may prefer (the revised) D1 over D3 and others (the revised) D3 over D1. (This possibility was illustrated in our discussion of the possibility of conflicts between the P-Pareto principle and the J-Pareto principle, even when our list of primary goods are extended to include leisure, on page 195.) Therefore, the two-step procedure and the rest of our argument will be applicable on this revised metric too.

7.8 Conclusion

Cohen's challenge to the Pareto argument for inequality was very simple yet very effective. Cohen argued that the Pareto principle did not justify departing from equality, because another equal distribution which was also Pareto efficient was feasible. In other words, Cohen argued that the move

from D1 to D2 wasn't justified by the Pareto principle, because D3, an equal and Pareto efficient distribution, was feasible. As a corollary of this, he argued that the talented should be committed to an egalitarian ethos that would bring about D3. In response, I distinguished between different Pareto principles. If Rawlsians needed to be committed to J-Pareto optimality, then, Cohen's argument would have force. However, I have argued that Rawlsians should be committed to P-Pareto optimality, which in Cohen's standard case only justifies a move from D1 to D2, and not a move from D1 to D3. Since the moral desirability of D3 has not been established, the case for an egalitarian ethos, which was recommended because it can bring about D3, also has not been established.

The weakest point of the argument of this chapter is the combination of primary goods as the metric of justice with P-Pareto optimality. In section 7.2 I offered a few reasons to mitigate the prima facie implausibility of this view. However, these reasons are far from conclusive. Problems with this view and J-Pareto optimality may be reasons for adopting welfare as the metric of justice. As we will see in the next chapter, the adoption of welfare as the metric of justice does not help the case for an egalitarian ethos.

Chapter 8

Egalitarian Ethi and Welfarism

8.1 Introduction

Let us begin by briefly recapitulating Cohen's argument. Cohen's argument was a response to the Pareto argument for inequality as it's presented by Barry. Cohen argued that when D2 is feasible, another distribution, D3, in which everyone is better off than they are in D1 and in which equality is preserved, could also be realized if the talented were to choose to act differently.[1] The three distributions were as follows:

	D1	D2	D3
Talented	d	a	b
Untalented	d	c	b

Table 8.1

[1]G. A Cohen, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008), p. 100-2.

Here letters represent each individuals' shares, and $a > b > c > d$.

In Cohen's *standard case*, which is the case Cohen identifies as the focus of his discussion, the talented prefer D2 over D1 and D1 over D3.[2] It is because the talented prefer D1 over D3, and act in accordance with this preference that the conflict between equality and Pareto-efficiency arises. Cohen argues that the attitude of the talented, which makes inequalities necessary, is in conflict with their supposed affirmation of the first stage of the argument presented by Barry. If the talented were motivated by an egalitarian ethos in their occupational choices, which includes how much to work and in which occupations to work, then D3 could be realized. Cohen, in effect, offers a reconciliation of equality and Pareto-optimality. The talented are as productive as they are in D2 without requiring inequality-generating rewards, because they act in accordance with an egalitarian ethos. This way we have both equality and Pareto-efficiency.

In the previous chapter, I offered a reinterpretation of the Pareto argument for inequality that undermines the case for an egalitarian ethos. The argument assumed that the primary goods were the correct metric of justice. In this chapter, I drop this assumption. Here, I depart from the Rawlsian framework, and evaluate the case for an egalitarian ethos on the assumption that the metric of justice is welfare, and D1 and D3 represent equal divisions of welfare.

[2]Cohen, *Rescuing Justice*, pp. 57ff.

The question of whether there is a case for an egalitarian ethos when welfare is the metric of justice is of interest in its own right. In addition to this, it is of interest, because the move to a welfarist egalitarianism is a way to fend off some of the objections to an egalitarian ethos. For instance, Cohen argues that his proposal for an egalitarian ethos avoids the charge that it would amount to the 'slavery of the talented', because he employs a 'broadly welfarist framework'.[3] Accordingly, by evaluating the case for an egalitarian ethos when welfare is the metric of justice, we will be evaluating its most defensible form.

I consider the case for an egalitarian ethos for three different accounts of welfare, which I define in the next section. I argue that if we assume that actual preference satisfaction is the correct theory of welfare, and we are in a position to bring about equality of actual preference satisfaction, then the conflict between equality and Pareto-efficiency, which Cohen offers the egalitarian ethos as a solution to, does not exist. Everyone, including the talented, would prefer D3 to D1. We would be in a position to bring about Pareto optimality in the space of actual preference satisfaction without having to rely on an egalitarian ethos. The argument for this claim, and the responses to possible objections take up sections 8.3 to 8.6.

In the second half of this chapter, I consider the case for an egalitarian ethos when either the objective list theory or the informed preference satisfaction theory gives the correct account of welfare. I first examine whether market-

[3]Cohen, *Rescuing Justice*, p. 370.

based egalitarian ethi, such as the ones offered by Joseph Carens and Martin Wilkinson, can bring about Pareto optimality in the space of welfare when welfare is so understood (section 8.7).[4] I argue that market-based egalitarian ethi can't bring about this goal unless people's actual preferences correspond to what these theories of welfare would recommend for them. When, however, people's actual preferences conform to the prescriptions of the correct theory of welfare, we're back to pursuing equality and Pareto optimality in the space of actual preference satisfaction. Consequently, the result established in the previous sections applies and the egalitarian ethos is redundant.

In section 8.8, I consider a hypothetical egalitarian ethos which brings about Pareto optimality in the space of welfare –when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory– even if people's actual preferences differ from what the correct theory of welfare prescribes for them. I argue that this hypothetical egalitarian ethos would be objectionably paternalistic. It would require bringing about outcomes against the wishes of individuals. This argument completes my case that when the metric of justice is welfare, an egalitarian ethos is either redundant or objectionably paternalistic.

[4]Joseph H. Carens, *Equality, Moral Incentives, and the Market: An Essay in Utopian Politico-Economic Theory*, (Chicago: University of Chicago Press, 1981); T. M. Wilkinson, *Freedom, Efficiency and Equality*, (Basingstoke: Macmillan, 2000).

8.2 Different theories of welfare defined

In this section, I define the different accounts of welfare I refer to in this chapter.

According to *actual preference satisfaction theories of welfare*, a person's life goes well to the extent that her actual preferences are satisfied. I assume that preferences satisfy the properties required for the first and second fundamental theorems of welfare economics. In the rest of this chapter, I shall refer to actual preference satisfaction as *utility*.

According to *informed preference satisfaction theories of welfare*, our lives go better when the preferences we would have if we were fully rational, well-informed, and free from distorting influences are satisfied. The condition 'free from distorting influences' influences is intended to rule out cases of morally objectionable adaptive preference formation where someone's preferences adapt to their unfairly imposed conditions.[5]

According to *objective list theories of welfare*, there are certain goods, such as friendship, freedom from pain, accomplishment etc. that make a life go better. Which goods make it on the list is independent of the beliefs of individuals. However, the goods on the list may refer to the mental states of people. So pleasure may be on the list, and, in that case, what makes

[5]There are doubts about the possibility of defining distorting influences without (a) relying on a notion of legitimate expectations which would be a problem for welfarist egalitarianism as a theory of distributive justice; and (b) referring to an objective account of welfare, which would be a problem for informed preference satisfaction theories. I set both of these problems aside.

Person A's life go better would have to refer to what gives Person A pleasure. However, pleasure would be on the list even if Person A did not think that pleasure contributed to a person's welfare.

8.3 The general argument for the redundancy of egalitarian ethi

With these definitions in place, we can now offer an intuitive argument for why an egalitarian ethos would be redundant when we are pursuing equality and Pareto-optimality in the space of utility. As you will recall, in Cohen's standard case the talented prefer D2 over D1 and D1 over D3. Given a choice between D1 and D3 they would choose D1 even though in D3 they receive more of the metric of justice.[6] This is why they need to be offered inequality-generating rewards to be more productive. Now, suppose that the metric of justice is utility. In that case, Cohen's standard case cannot arise. For by definition, the talented prefer D3 over D1, because D3 contains more of the metric of justice, which we are assuming is utility. So if the metric of justice is utility, the standard case does not arise. In other words, we will not have the conflict between equality and Pareto-efficiency which recommended D2. We will be in a position to realize D3. This is bad news for egalitarian ethi which are put forward as proposals to help us overcome the conflict between equality and Pareto-efficiency. The absence of a conflict between Pareto-efficiency and equality makes the egalitarian ethos redundant.

[6]This is what distinguishes the standard case from the bad case. In Cohen's *bad case*, the talented actually prefer D3 to D1, but pretend otherwise for bargaining purposes. Cohen, *Rescuing Justice*, p. 102.

This argument involves a big 'if' which I should bring to the fore. It assumes that we are in a position to bring about equality of utility. What entitles us to this assumption? Suppose we cannot bring about equality of utility, and utility is the metric of justice. Instead, we bring about equality of income, and there is an egalitarian ethos in place. One objection to this scheme would be that it would leave some talented people worse off than the untalented, because it does not compensate them for their labour burden. This scheme would not be an improvement from the perspective of justice. A second objection is that if we only have equality of income, we haven't made progress in our response to the dilemma that the egalitarian ethos is offered as a solution to. We don't have equality in the space that justice requires equality in.

Let me note that Cohen himself thinks that the egalitarian ethos should be instituted within a welfarist perspective. One objection to instituting an egalitarian ethos and equality of income would be that this scheme condemns those talented people, who hate the jobs they are most productive at, to a life they find oppressive. Cohen acknowledges the force of this objection. To fend off this objection to the egalitarian ethos, the duty Cohen proposes caters to the subjective burdens and benefits of different occupations. He requires that 'no one should demand a salary much more than what is required for her to be roughly on a par with others, given the satisfactions, and frustrations, that attach, for her, to her job'.[7] So, Cohen grants that, in the absence of equality in a metric which caters to the subjective burdens and benefits of

[7]Cohen, *Rescuing Justice*, p. 370.

occupations, an egalitarian ethos would have objectionable consequences.

Our argument presented in the beginning of this section does not need to assume that we are in a position to bring about equality of utility in a precise fashion. Cohen, realistically, intends the duty and the equal distribution to be roughly applied. What Cohen says in his camping trip example illustrates Cohen's intent: we are required to contribute 'roughly equally and enjoy the fruits of our cooperation roughly equally'.[8] This assumption of rough equality of utility is enough to run the same argument. Whether roughly or precisely equal D3 will provide more utility than D1.[9] So, everyone will prefer D3 over D1.

Even if all this is correct, isn't an egalitarian ethos necessary for achieving *equality* of utility? What is required is not an egalitarian ethos, but compliance with the demands of egalitarian institutions. The egalitarian ethos as Wilkinson and Carens present it is not intended to bring about a certain distribution, but to ensure that the amount of goods available for distribution is not influenced by the principle of distribution in place. Egalitarian ethi like those offered by Carens and Wilkinson are intended to 'break the link between production and distribution'.[10] Wilkinson elaborates this point:

> 'The incentives objection, that equality must conflict with freedom or efficiency, rests on the claim that production is not inde-

[8]Cohen, *Rescuing Justice*, p. 352.

[9]This is true provided that the gap between D1 and D3 is larger than the variations allowed for a distribution to count as equal.

[10]Wilkinson, *Freedom, Efficiency and Equality*, p. 126. Carens also states that this is the aim of his social duty. See Carens, *Equality, Moral Incentives, and the Market*, p. 183.

> pendent of distribution, that what people would produce depends on what is produced. If a satisfactory scheme of social duty can be worked out, then, at least in principle, production and distribution can be separated.'[11]

The social duty breaks the link between production and distribution because people, informed by the egalitarian ethos, produce not with an eye to what they will end up with once everything gets distributed equally, but in accordance with a social duty. People produce according to the social duty, and what's produced gets distributed according to the correct conception of justice.

The egalitarian ethi offered by Carens and Wilkinson are *not* intended to bring about a certain distribution. The distribution is brought about by other means. On this point, Wilkinson's discussion is instructive. He writes,

> 'If the social duty achieves the aim of efficient production, then the question of how the fruits of that production should be distributed can be left to one side. Those fruits could be distributed in whatever is the right way. Generally any claim that social duty is unfair falls before this point: one cannot press an unfairness objection to any social duty without presupposing some kind of metric to measure unfairness and, once one has found that metric, people could be given the fair shares it recommends.'[12]

[11] Wilkinson, *Freedom, Efficiency and Equality*, p. 126.
[12] Wilkinson, *Freedom, Efficiency and Equality*, p. 145.

Carens and Wilkinson are correct in not claiming that the egalitarian ethos can ensure distributional equality. Individuals are in a better position than social institutions to know their productive capacities and utility functions. Egalitarian ethi utilize the first aspect of individuals' comparative advantage, their better knowledge of their productive capacities. Despite these advantages, however, individuals are not in a better position than social institutions to know how their condition compares with that of others. I can know how much of my productive capacities I'm using, but I can't make sure that I'm not worse off or better off than others. I can also imagine that I prefer my income and job-satisfaction package to those of many others, but that is not a relevant comparison. All that this comparison would show is that I'm better off than I would be if I had that person's income and job. The relevant comparison, however, is whether I enjoy more utility than that other person. To make this comparison, one needs to know how much utility they get from their incomes and occupations. This requires knowledge of others' occupations, income, and utility functions. Clearly, social institutions are better placed to acquire this information.

Given the fact that institutions are in a better position to bring about (rough) equality of utility than individuals, the task of bringing about a (rough) equality of utility is one which should be left to social institutions. For the social institutions to bring about (rough) equality of utility, I will be under a duty to provide information about my utility function and the goods I enjoy, but this is a duty distinct from the duty required by the egalitarian ethos. It does not go beyond the duty to comply with the laws

of just institutions. It is simply the truthful revelation of information about oneself and one's resources that is necessary for taxation purposes.

As we've seen Cohen, in his formulation of the egalitarian duty, explicitly specifies that the egalitarian ethos does not require the talented to be worse off than others. However, he is not very clear about how he thinks equality ought to be achieved. I have argued here that egalitarian ethi offered by Carens and Wilkinson are not intended to bring about a certain distribution and that an equal distribution of utility can't be brought about by individual efforts, but would require institutional efforts. I shall return to Cohen's position on whether the egalitarian ethos has a distributive role in section 8.6.

The argument of this section can be summarized as follows. Either (rough) equality of utility can be brought about, in which case the egalitarian ethos is redundant, because the talented would also prefer the outcome in which they are more productive, and, accordingly, there's no conflict between equality and Pareto-efficiency, or (rough) equality of utility cannot be brought about, in which case the egalitarian ethos is morally unjustified.

8.4 Optimal income taxation and the theory of incentives

In this and the next section, I develop a more detailed account of the claim that if we are in a position to realize equality of utility, we will also be in

a position to realize Pareto optimality in the space of utility. Showing this requires a brief detour into the literature on optimal income taxation and the theory of incentives.

The second fundamental theorem of welfare economics is the starting point of the literature on optimal taxation. According to the second fundamental theorem of welfare economics, if a social planner knows the utility function, which satisfies certain formal conditions, and endowments of each and every individual in society, then the full optimum allocation of goods given by a welfare function can be realized by the social planner through the following scheme: (a) the social planner announces a set of prices; (b) the social planner rearranges initial endowments, which include talents, through lump-sum taxes and subsidies; and (c) the social planner allows individuals to trade at these prices.[13] In addition to the utility functions and the endowments of individuals, the social planner would have to know how hard individuals work and localised pieces of information such as technological possibilities. Following the redistribution of endowments and the announcements of prices, individual trading at these prices and staying within their budget constraints will trade up to the specified social welfare function.

The theorem relies on lump-sum transfers which are implemented independently of the behaviours of individuals. For this reason the theorem requires extensive information about individuals. It is generally conceded that in

[13]I am here following Dasgupta's very accessible account in 'Utilitarianism, Information and Rights'. See Partha Dasgupta, 'Utilitarianism, Information and Rights', in: Amartya Sen and Bernard Williams, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982).

practice this information is hard to acquire, because individuals will have incentives to withhold information about themselves. As a result we have an incentive problem. For instance, the realisation of the utilitarian optimum would require that the talented work extra hours in order to increase the sum of utilities, which in turn results in their enjoying less utility than others. Accordingly, in a utilitarian society the talented will have incentives to hide their talents to avoid high tax rates.

According to the *principal-agent model*, asymmetries of information are the source of incentive problems.[14] On the principal-agent model there is a *principal* who delegates a task to the *agent*. The principal may lack information about the agent's characteristics, or about the agent's actions. This asymmetry of information enables the agent to maximize her own welfare at a cost to the principal. If the agent's advantage is due to the principal's lack of information regarding the agent's characteristics, we have a *hidden characteristic*, also called adverse selection, problem. For instance, if the social planner doesn't know the skill levels of individuals we will have a hidden characteristic problem. When the agent's advantage is due to the principal's lack of information regarding the agent's actions, we have a *hidden action*, also called moral hazard, problem. If, for instance, the social planner doesn't know the number of hours individuals work, we would have a hidden action problem. In both cases the principal needs to devise an incentive structure so that the agent acts in a way that maximally benefits the principal. In our

[14]For overviews of the principal-agent model see Donald E. Campbell, *Incentives: Motivation and the Economics of Information*, (Cambridge: Cambridge University Press, 1995); and Jean-Jacques Laffont and David Martimort, *The Theory of Incentives: The Principal-Agent Model*, (Princeton, N.J: Princeton University Press, 2002).

case, the social planner is the principal and the individuals being taxed are the agents. The relevant information is the endowments, including talents, of individuals, their utility functions, and how much they work. The problem of incentives arises because the principal lacks information about the agent and the agent can use this fact to improve his condition.

8.5 Dasgupta and Hammond on incentive-compatibility and equality

Since Mirrlees's pioneering work, economists have been exploring how to implement optimal taxation given informational constraints the social planner faces.[15] Dasgupta and Hammond's work shows that, under certain conditions, an equal division of utility is both Pareto-efficient and incentive-compatible even when all the information demanded by the second fundamental theorem of welfare economics is not available.[16]

Dasgupta and Hammond employ a model which is a variation on Mirrlees' model. On Mirrlees' model individuals have productive capacities which they can turn into a single consumption good that is called income. Each individual's utility is a function of their net income and the amount of labour they supply. Everyone is assumed to have the same utility function, but

[15]J. A. Mirrlees, 'An Exploration in the Theory of Optimum Income Taxation', *The Review of Economic Studies*, 38 (1971):2. For an accessible presentation of Mirrlees' model, see Matti Tuomala, *Optimal Income Tax and Redistribution*, (Oxford: Clarendon Press, 1990).

[16]Partha Dasgupta and Peter Hammond, 'Fully Progressive Taxation', *Journal of Public Economics*, 13 April (1980):2; Dasgupta, 'Utilitarianism, Information and Rights'.

have different skill levels, which are defined as capacities to turn labour into income. Individuals decide how much labour to supply in light of what will maximize their utility. By altering the informational constraints the social planner is under, we can work out how to implement a given social welfare function, or a second-best.

Dasgupta and Hammond's model examines whether a social planner can implement a first-best utilitarian optimum. They assume that the social planner knows everyone's utility function, which is assumed to be identical for everyone, and that leisure is a normal good. It is also assumed that the social planner knows the number of hours a worker works and his income. Under these conditions, if the worker works at his true skill level, then, 'it is possible to infer the worker's skill, since it is (in Mirrlees' model) just proportional to the income he earns per hour of work'.[17] In this case, a first-best utilitarian optimum cannot be implemented, because, as we have seen, it requires the more skilled workers to end up with lower utility. For this reason, the skilled workers would choose not to reveal their true skill levels. Dasgupta and Hammond conclude that 'the best we can do without knowing workers' skills in advance is to devise a taxation scheme which leaves all workers equally well off'.[18] Since those workers who have higher skill levels won't be worse off than others, they will not have an incentive to hide their skills. We can therefore implement lump-sum taxes based on revealed skills. This outcome will also be Pareto-efficient.

[17]Dasgupta and Hammond, 'Fully Progressive Taxation', p. 142.
[18]Dasgupta and Hammond, 'Fully Progressive Taxation', p. 142-3.

Dasgupta and Hammond's model assumes that leisure and income are the only sources of utility. However, their model can be extended to include other sources of utility or disutility such as the effort expended while carrying out a task or the appeal an occupation might have for a person. Their assumption that everyone has the same utility function can also be dropped. The key point of their analysis is that if the social planner is in a position to ensure equality of utility, then he will be in a position to bring about an equal and Pareto-optimal distribution. So, when we introduce the complications I've mentioned into their model, we need to assume that the social planner knows how much of these other sources of utility a person has, and what their utility function is.

We can put it this way: A social planner who is in a position to bring about an equal distribution of utility faces only a hidden characteristic problem, and doesn't face a hidden action problem in bringing about the preferred outcome. This is because all of the relevant actions, such as how many hours a person works, how much effort they exert etc. would have to be known to bring about an equal distribution of utility.[19] And given this information, taxation which leaves everyone equally well off will be incentive compatible, because no one will have an incentive to hide their talents.

The informational requirement in Dasgupta and Hammond's scheme is tremendous. The social planner would have to know everyone's utility functions, how much they work, and the distribution of productive capacities

[19]I'm grateful to Jerry Cohen for impressing on me the need to clarify this point.

within a society. However, it doesn't compare badly with a scheme ensuring *only* an equal distribution of utility. The only piece of extra information that is needed is the distribution of productive capacities within the society, which is not a significant informational burden. (The social planner would already need to know how much everyone works since work is a source of utility or disutility as the case may be.) If we are in a position to ensure an equal distribution of utility, then we will also be in a position to realize a Pareto-efficient outcome through lump-sum taxes. The result established by Dasgupta and Hammond shows that for an equal and efficient distribution of utilities to be achieved the individuals do not need to inform the social planner what their productive capacities are. They will reveal their skill levels indirectly, not because they are bound by a moral duty to reveal it, but because revealing their skill levels won't make them worse-off.

It may be thought that Dasgupta and Hammond's scheme relies on an egalitarian ethos, and requires not only that people comply with the demands of justice governing what Rawls calls the basic structure of society, but also that they comply with a moral duty in their interactions outside the basic structure. Let us assume a particularly narrow definition of the basic structure, and define it as the coercive structure of society. Even on this narrow definition, complying with the demands of the tax system falls within the basic structure, and this is all we need for Dasgupta and Hammond's scheme. In Rawls's theory, people are assumed to act as self-interested maximizers in the market and comply with the requirements of the tax system. For instance, they will reveal their earnings truthfully for taxation purposes,

or if there are benefits paid to people with disabilities, they will not lie to receive these benefits they are not entitled to. For a scheme which ensures an equal distribution of utility, they will truthfully reveal their income and utility function. The government will, then, impose lump-sum taxes based on this information and the information on their skills which it has inferred from their income and the number of hours they have worked. In their market transactions, people will still be acting as self-interested maximizers. They will not be required to comply with a duty that goes beyond complying with the demands of the tax system.

8.6 The liberty objection to lump sum taxes

Aren't the egalitarian ethi preferable to Dasgupta and Hammond's scheme, because they, unlike lump sum taxes, do not infringe on liberty? If we do not think that equality of utility conflicts with liberty, then we are unlikely to find lump-sum taxes to be objectionable on grounds of liberty. To see this, consider a hypothetical scheme that is similar to Dasgupta and Hammond's but doesn't rely on lump-sum taxes on talents. Suppose we have a two person economy where one worker is more skilled than the other. The social planner knows each individual's utility function, and the distribution of talents, but not who possesses these talents. The social planner announces that she will ensure an equal distribution of utility, and informs both workers of the possible outcomes their occupational choices can bring about. The social planner also informs them that equality of utility at the highest feasible level would be achieved if the talented worker, whoever she is, worked in Job A. Given this

piece of information, the talented worker would choose Job A. The possible outcomes that the agents would need to know about in order to make their choices increases exponentially with the number of people in the economy. So this is not a remotely practical proposal. Nevertheless, it is analogous to the proposal made by Hammond and Dasgupta. In their proposal, the social planner imposes lump-sum taxes and the agents coordinate their activities in the market. In the hypothetical scheme, the social planner provides the agents with the information they need to sort out the coordination problem without the need of the market. If we do not find this hypothetical scheme, and the realization of equality of utility, objectionable, then it is hard to see why we should find the imposition of lump-sum taxes on talents objectionable. In Dasgupta and Hammond's scheme, lump-sum taxes are introduced to help people realize an outcome that they would prefer, but cannot realize due to difficulties of coordination. Our judgements about whether lump-sum taxes infringe on liberties in a morally objectionable way should be the same as our judgements in the hypothetical scheme.

We can use an example from Cohen to illustrate our response to the liberty objection and our argument that when we can have equality of utility, the egalitarian ethos is redundant. Cohen gives the example of a person, A, who has the following preference orderings.

> 'a. A doctors at £50,000: she is much better off than most people are in both job satisfaction and income
> b. A gardens at £20,000: she is much better off than most people are in job satisfaction (even more than she was in a.) but not

better off in income

c. A doctors at £20,000: she is still much better off than most in job satisfaction but she is not better off in income.'[20]

Cohen also stipulates that the community's preference ordering is c>a>b. Cohen then states the trilemma, which he claims the egalitarian ethos resolves, thus.

'The trilemma that the egalitarian is said to face runs as follows. If, in deference to equality and freedom, we freeze salaries at £20,000 and allow the doctor-gardener to choose her work, she will garden, and then both she and the rest of the community will be worse off than they could be: Pareto will be violated. With pay equal, and freedom of choice of occupation, we get the Pareto-disaster that consumers have no say in what gets produced (in this case, some package of gardening and doctoring services). But, if in deference to freedom and Pareto, we offer the doctor-gardener £50,000 for doctoring, then equality goes. And if, finally, in deference to equality and Pareto, she is forced to doctor at £20,000, then freedom of occupational choice is lost.'[21]

What Cohen says here is not wrong, but by focusing on income equality he's focusing on the wrong kind of equality. Assuming that income and job satisfaction are the only goods in play and everyone's utility function for income is the same, when A gardens and ends up with £20,000, she's still

[20]Cohen, *Rescuing Justice*, p. 185.
[21]Cohen, *Rescuing Justice*, p. 185-6.

better off than other people. She has the same income as others and more job satisfaction. So, the outcome in which A gardens and ends up with £20,000 is not an equal outcome in the relevant sense. If we are pursuing equality of utility, then A should be more heavily taxed when she gardens, and that extra taxation should be given to others to ensure equality of utility. Let's say, hypothetically, that equality of utility is achieved when she receives £15,000 and gardens, or when she doctors and receives £18,000, and the latter equal distribution is the Pareto-efficient one. In this case, she should face a tax rate that gives her those two options.[22] And when this tax rate is in place, A will choose doctoring, because it contains more utility.[23] Setting up a higher tax rate for A when she gardens in order to ensure equality of utility is not, intuitively, an infringement of her liberties. She can still choose to garden. However, she will not, because doctoring and earning a higher income is preferable.

It's very difficult to see why taxing A more heavily when she gardens should be objectionable on grounds of liberty unless the pursuit of equality of utility was objectionable in the first place. Cohen suggests one reason which may be relevant.

> '... [W]e cannot tell what people's total situations are, exactly, and there is consequently a danger of coercing those whom we wouldn't coerce into a particular job (even if anybody should ever

[22]She should receive less than £20,000 even when she doctors, because Cohen stipulates that she is better off than most in job-satisfaction, and we assumed that everyone had the same utility function for income.

[23]Since both distributions are equal distributions of utility, and one of them is Pareto-efficient while the other is not, it contains more utility for everyone.

> be coerced into a job) if we had knowledge about them that we can't have. I believe that the rough-and-ready "everyone must do his or her bit" principle that prevailed in the Britain of World War Two was a principle of justice, despite an inevitable uncertainty as to whom it commended and condemned, but it would have been grotesque to try to enforce it coercively, because one person's easy bit is another person's hard bit, and figuring out what's hard for whom is an unmanageable task. We similarly can't tell how much the doctor-gardener dislikes doctoring, or not without an enormously invasive apparatus. (It does not follow that she is as much at a loss as we are with respect to knowing the size of that dislike, and therefore, whether her salary demand is defensible from the point of view of egalitarian justice.)'[24]

Several points need to be against employing this line of argument against schemes which bring about equality of utility through institutional measures, such as the lump sum taxes offered by Dasgupta and Hammond. Firstly, it is true that the doctor-gardener is better positioned to know the extent of her dislike. I granted this point in section 8.3. However, it does not follow from this that she is also in a better position to know whether 'her salary demand is defensible from the point of view of egalitarian justice'. She is not in a good position to tell how her condition compares to that of others. She may know that doctoring at £20,000 is not unbearable for her, but she can't know whether she is better off or worse off than others.

[24]Cohen, *Rescuing Justice*, p. 219. Cohen here is not discussing lump-sum taxes, but conscription.

Suppose Cohen's proposal is to have equality of income through taxation, and have people follow the egalitarian ethos with the proviso that they do not end up worse off than others. So, the doctor-gardener should choose to doctor at £20,000 if she believes she won't be worse off than others. Given my first point that people can't know whether they are worse or better off than others, the trilemma, Cohen has described, has not been dissolved. We don't have equality in the space that justice demands equality in. We only have equality of income. (On this construal of his proposal, Cohen is out of step with Wilkinson and Carens who, rightly, don't assign distributive aims to their proposed egalitarian ethi.)

Furthermore, institutional measures aimed to bring about equality of utility need not be invasive. We can make rough and ready judgements about different people's utilities. These rough and ready judgements need not be invasive, particularly when we assume that people are committed to the principles of justice in question. It is precise judgements which would be invasive. We can make generalizations about which occupations are particularly enjoyable, and which ones are particularly arduous. It is true that some of these judgements would be mistaken. As a result, our tax rate would make some people worse off than others, and result in some efficiency losses. However, the extent of these inequalities and efficiency losses would be less than the proposal we looked at in the previous paragraph, because, as we have observed, the government is better placed than individuals to make these judgements.

One last point. Suppose we grant Cohen that the doctor-gardener can make a rough judgement about how she compares with others. The doctor-gardener decides whether to doctor or not based on her own judgement of whether doing so would or would not make her worse off than others. In this case, Cohen has to explain how a person can know how her condition compares with that of others, and why that information would not be available to those who are setting the tax rates.

Before moving on to the discussion of egalitarian ethi when welfare is different from actual preference satisfaction, I would like to emphasize a point about Dasgupta and Hammond's scheme to avoid misunderstandings about the relationship between the argument of section 8.3, and the arguments of sections 8.4 to 8.6. On Dasgupta and Hammond's scheme, lump-sum taxes are not offered as a measure over and above those measures necessary to bring about equality of utility. Rather, they are the most natural way of bringing about equality of utility in a complex economy. (The scheme I proposed on page 247 would be impossible to carry out in a complex economy.) It is a separate fact, which is in line with the argument of section 8.3, that such lump-sum taxes will produce an outcome that is also Pareto-optimal. So, the argument is not that when we are in a position to bring about equality of utility, we can also make sure that this outcome is Pareto-optimal by relying on lump-sum taxes. The argument is that the pursuit of equality of utility, which would have to be carried out through a scheme like Dasgupta and Hammond's, would also produce a Pareto-optimal outcome.

By explaining how I see the dialectic as unfolding, I can further clarify what I consider to be the force of the redundancy objection and Dasgupta and Hammond's proposal. The egalitarian ethos is offered as a solution to an incentive problem. The idea is that we can either have equality or Pareto-optimality. The egalitarian ethos is offered as a way to overcome this dilemma. The critic of the egalitarian ethos objects that the egalitarian ethos would be unfair to the talented. Cohen suggests that it wouldn't be, because the equality in question is (something like) equality of utility. But when we're in a position to have rough equality of utility, which requires that institutions have access to information about individuals' utility functions and labour burdens that normally only individuals have access to, the conditions that gave rise to the incentives problem, i.e. the problem of motivation and the asymmetry of information, no longer exist. Therefore, the egalitarian ethos is redundant. So, I'm not offering Dasgupta and Hammond's proposal, because I think it is a more effective solution to the trilemma that exercises Cohen. It is clear that the information in their scheme is very difficult to come by. It only shows, starkly, that under the conditions when the egalitarian ethos is desirable, there's no need for it. Furthermore, if my argument about the ineffectiveness of an egalitarian ethos to bring about equality of utility is correct, the egalitarian ethos can't bring about the conditions under which it would be desirable. The egalitarian ethos is not a solution to the dilemma between equality and efficiency, because when we can have equality of utility, we will also have efficiency. And, when we can't have equality, an egalitarian ethos can't help us achieve equality. Speaking colourfully, we might say, the egalitarian ethos is like a cold medicine that is only safe to take when one

does not have a runny nose and a sore throat.

8.7 Market-based egalitarian ethi when welfare isn't actual preference satisfaction

We have seen that if we are in a position to bring about equality of utility, we can bring about Pareto-optimality in the space of utility without relying on an egalitarian ethos. So, for welfarist egalitarians who hold that welfare is utility, there isn't a case in favour of an egalitarian ethos. In the rest of this chapter, I consider the case for an egalitarian ethos when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory. In this section, I examine whether the market-based egalitarian ethi offered by Carens and Wilkinson, and broadly affirmed by Cohen, can bring about Pareto optimality in the space of welfare – when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory.

The egalitarian ethi offered by Carens and Wilkinson are intended to bring about equality and Pareto-efficiency by relying on the market mechanism. They involve (a) 100% taxation on income and its redistribution in accordance with the principles of justice, (b) a moral duty that individuals are required to follow which governs their occupational decisions such as how much to work and which occupations to take; and (c) the use of markets.[25]

[25]Carens, *Equality, Moral Incentives, and the Market*; Wilkinson, *Freedom, Efficiency and Equality*.

The duty proposed by Wilkinson is for individuals to 'respond to market prices in the egalitarian system as though... [they] were getting the money for... [their] own personal consumption'.[26] According to Carens's most recent formulation of his egalitarian duty, people are not required to maximize their pre-tax incomes, but 'to work or more or less full time... and to make good use of their talents and skills' despite the 100% tax rate.[27]

Proposals like those offered by Carens and Wilkinson, which employ the market, cannot promote Pareto-efficiency in all spaces. The first fundamental theorem of welfare economics, which shows the Pareto-optimality of markets under certain conditions, works only for Pareto-optimality in the space of utility. When people engage in trade or production their choices are guided by their actual preferences. When I express a demand in the market, it is a reflection of my actual preference. I cannot act on the preferences I would have if I had deliberated with full information unless I've already done so and my actual and informed preferences correspond. For this reason, markets need not be Pareto-optimal when welfare is construed as being different from utility.[28]

[26]Wilkinson, *Freedom, Efficiency and Equality*, p. 136.

[27]Joseph H. Carens, 'Rights and Duties in an Egalitarian Society', *Political Theory*, 14 (1986):1, p. 35.

[28]Sen's extension of the efficiency of competitive market equilibria to weak-efficiency in terms of what he calls opportunity freedoms may be considered to be a counter-example to my claim. However, his definition of freedom involves preferences. This is the result of his Axiom O which reads 'To be sure that A offers more opportunity-freedom than B (alternatively, at least as much as B), there must be an element of A that is preferred to (alternatively, regarded as at least as good as) all the elements of B.' See Amartya Sen, 'Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms', *Oxford Economic Papers*, 45 October (1993):4, p. 530.

Let me illustrate this problem with an example. Suppose the correct theory of welfare is informed preference satisfaction, and informed preferences exclude some adaptive preferences. Someone motivated by an ethos like Carens' or Wilkinson's will have to see the satisfaction of the actual preferences of others as part of his duty even though this may not be what their welfare consists of. Consider this case.

> *The writer* I'm a writer in the 1960's. I face two options: I can either write a book that will contribute to women's liberation, but will not sell much, or I can write a book that will reinforce the preferences of housewives whose preferences have adapted to their condition that is unfairly imposed on them. I know that this second book is going to be a best-seller.

Under the egalitarian ethos as specified by Carens or Wilkinson I will have a *moral* duty to take the second course of action. However, if welfare is informed preference satisfaction writing the best-seller would bring harm rather than benefit.

There are two possible responses to the fact that markets only produce Pareto-optimality in the space of actual preference satisfaction. Firstly, it may be suggested that we should have mechanisms that will ensure that people's actual preferences will correspond to their informed preferences or with the recommendations the objective list theory would offer them. It is difficult to spell out how this goal might be achieved. The following measures are a first approximate: providing people with the education that will enable them to reason well about their welfare and make information which

is relevant to their interests available, and offering psychological help, such as Brandt's 'cognitive psychotherapy', to remedy problematic instances of adaptive preference formation.[29] This is far from a complete theory of how people's actual preferences may be made to correspond with their informed preferences or with what the objective list theory of welfare would recommend. Nevertheless, it gives some idea about the most effective means to pursue it. What seems clear is that these goals would most effectively be pursued by institutional measures.

Suppose we have these measures, which will ensure that people's actual preferences correspond to what the correct theory of welfare would prescribe to them, in place. People's actual preferences, to a large extent, correspond to what they ought to prefer according to the correct theory of welfare. Equality of actual preference satisfaction will entail equality of informed preference satisfaction or equality of welfare according to the objective list theory. In that case, however, the results of section 8.3 will apply. We will be pursuing equality of utility and Pareto-optimality in the space of utility. Accordingly, an egalitarian ethos which governs people's occupational decisions will be redundant.

The second response to the fact that markets produce Pareto-optimality only in the space of actual preference satisfaction is to adopt actual preference satisfaction as a second best. We can reason as follows. Surely, people's actual preferences do not correspond to what they would prefer if they were

[29]Richard B. Brandt, *A theory of the Good and the Right*, Rev. ed. edition. (Amherst, N.Y.: Prometheus Books, 1998).

fully informed, or what the objective list theory of welfare would recommend. But, every individual is different and they would prefer different things if fully informed. A similar point applies to some objective list theories of welfare. There is an objective list that gives the correct account of welfare for an individual, but the list need not be the same for all individuals. The objective list theory may, for instance, include accomplishments. But people differ in the areas in which they can have accomplishments. So, the theory's recommendations for different individuals will be different. Given the role of facts about the individual in determining what the objective list theory or the informed preference satisfaction theory requires in their case, individuals themselves are the ones most likely to approximate the recommendations of these theories. Their actual preferences may be mistaken here and there. Yet, in most cases, actual preferences of individuals are more likely to approximate what these other theories of welfare would recommend for them than other alternatives. So, we should pursue Pareto-optimality in the space of utility as a second-best. This same line of reasoning would apply to the space in which we seek equality. Accordingly, the pursuit of equality of utility would be our second-best metric for equality. Now, we're back to pursuing equality of utility and Pareto-optimality in the space of utility. Accordingly, the results of section 8.3 apply. Again, the egalitarian ethos will be redundant.

8.8 The paternalism objection

Suppose we have developed an ethos which can bring about Pareto-optimality in the space of welfare – when the correct theory of welfare is either the in-

formed preference satisfaction theory or the objective list theory. Suppose that this hypothetical ethos produces this result without making the actual preferences of individuals conform to what the correct theory of welfare prescribes for them. (I must admit that I think it highly unlikely that such an ethos can be devised. Nevertheless, I examine its implications for the sake of completeness.) This hypothetical ethos can avoid the redundancy objection posed in the previous section. However, it faces another objection. It would have paternalistic consequences. By claiming that the ethos in question would have paternalistic consequences, I don't mean to suggest that it would be coercive.[30] On that point, I wish to remain neutral. What I mean is that the ethos would require people to bring about outcomes which they do not want, or want less than other feasible alternatives. More specifically, the ethos would require us to make certain options more costly or difficult for others to attain so that they pick options, which they otherwise would not pick, because the options they are made to pick better promote their welfare.[31]

Let me first illustrate how we can have someone acting paternalistically without coercing another person. Suppose, the newsagent on my street is the only place I can buy cigarettes from within the 25 mile radius of where I live. The newsagent thinks it would be better for me if I did not smoke. So, he will sell me a pack of cigarettes only for £50. I can either buy the pack for £50,

[30]For the view that not all instances of paternalism involve coercion see Gerald Dworkin, *The Theory and Practice of Autonomy*, (Cambridge: Cambridge University Press, 1988), p. 121-122.

[31]For the distinction between costly and difficult, see G. A Cohen, *Karl Marx's Theory of History: A Defence*, Expanded edition. (Oxford: Oxford University Press, 2000), p. 238-239.

in which case it will be too costly, or ride up to town to get a pack, which will be too difficult. In this case, the shopkeeper is acting paternalistically even though, intuitively, his action does not involve coercion.

Let us finally illustrate how an egalitarian ethos, which brings about Pareto-optimality in the space of welfare – when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory – without making the actual preferences of individuals conform to what the correct theory of welfare prescribes for them, would be paternalistic. Suppose we have achieved Pareto-optimality in this space through an egalitarian ethos. This entails that everyone in society is assigned a bundle of goods and an occupation necessary to achieve this outcome. They work in the jobs and consume the goods that would bring about Pareto-optimality in the space of welfare. *Ex hypothesi* there are other feasible outcomes that members of this society prefer to their current bundles and occupational choices, but they end up with this current bundle, because this bundle is the one that is good for them. This society would be intolerable, and rightly called paternalistic.

As just one example, consider the following case. Suppose everyone in my society prefers soda pops to *premier cru* clarets when, if well informed, we would prefer *premier cru* clarets to soda pops.[32] As it happens, I am the only person who can produce soda pops or *premier cru* clarets. Furthermore, my choice of occupation affects only what gets produced and nothing else. The ethi in question would require me to produce *premier cru* clarets instead

[32]This is a variation of an example offered by James Griffin. See James Griffin, 'Modern Utilitarianism', *Revue Internationale de Philosophie*, 141 (1982), p. 334.

of soda pops since the production of *premier cru* clarets would be a Pareto improvement in the space of informed preference satisfaction. In this case, the ethos brings about an outcome not preferred by anyone.

Admittedly, it is difficult to provide a detailed analysis of how and in what ways the hypothetical ethos would be paternalistic without a detailed account of the ethos. However, under the conditions we are assuming, the ethos will bring about results that are not wanted by its beneficiaries, and this is enough for the charge of paternalism.

8.9 Conclusion

Let me summarize the argument of this chapter.

1. If our theory of welfare is actual preference satisfaction, and if we are in a position to bring about equality of welfare, then we can bring about Pareto-efficiency in this space without relying on an egalitarian ethos. This is because, in the case of equality of utility, the conflict between Pareto-efficiency and equality, which the egalitarian ethos is offered as a solution to, does not exist. If we aren't in a position to bring about equality, an egalitarian ethos can't help us bring it about.

2. The egalitarian ethos is not preferable to a lump-sum tax on grounds of liberty.

3. On the assumption that the correct theory of welfare is not actual preference satisfaction, the market-based egalitarian ethi proposed by

Carens and Wilkinson cannot bring about Pareto-optimality in the space of welfare as long as people's actual preferences diverge from the prescriptions of the correct theory of welfare. This is because the market registers only the actual preferences of individuals.

4. Through various measures, people's actual preferences may come to conform to the recommendations of the informed preference satisfaction theory or the objective list theory. In that case, however, the result in (1) applies and egalitarian ethi are redundant. Similarly, if we rely on people's actual preferences as approximations of what the correct theory of welfare would recommend for them, then the result in (1) applies and egalitarian ethi are redundant.

5. If there is an egalitarian ethos, which brings about Pareto-optimality in the space of welfare – when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory – and people's actual preferences differ from the prescriptions of the correct theory of welfare, then this ethos would be objectionably paternalistic.

Chapter 9

Conclusion

9.1 Summary of the Arguments

Let's first summarize the ground we've covered by going over the main questions we addressed, and our answers to these question.

What is reflective equilibrium? Reflective equilibrium is an accurate description of the reasoning process of typical moral inquirers. It is possible that someone who has reached reflective equilibrium may have false beliefs. Depending on one's theory of justification, it is also possible that her beliefs are unjustified. Reflecting on the belief structure of typical moral inquirers reveals an important point about the priority of first-order moral claims over second-order ones. Insofar as metaethical theories conflict with first-order claims, they are unlikely to gain our acceptance. This is because the claims that support metaethical theories are usually ones that we are less confident of than our moral beliefs.

What is the role of other people's beliefs in our theorizing about justice? One uncontested role of other people is as facilitators in our thinking. The arguments they offer can help us see things we have reason to accept. Their beliefs can also play a role in our thinking *qua* the subjects of our moral concern. For instance, if we believe that we shouldn't impose principles that they do not have reason to accept, then we need to take into account their beliefs. We can offer reasons to them only on the basis of their beliefs. The mere fact that someone disagrees with us, by itself, however, is not always a reason for us to lower our confidence in the belief in question, or not to use that belief that they think is false as a premise in our thinking. One important result that follows from this is that what we are entitled to believe about a certain question and what we can expect others to come to believe on that same question can differ.

What is constructivism? Constructivism is not a single doctrine but several related doctrines. Bound constructivism offers a reduction of reasons, or values, within a given domain in terms of other reasons, or values. One is free to adopt different metaethical accounts of the reasons that form the reduction base in bound constructivism. Unbound constructivism offers an account of what it sees as the basic normative notion whether it is reason, or value *simpliciter* in terms of a viewpoint or procedure suitably constructed. Unbound constructivism, unlike bound constructivism, is a metaethical doctrine. Since bound constructivism can be combined with different metaethical doctrines, it can also be combined with unbound constructivism. Another variation amongst constructivist doctrines is due to their treatment of the

reasons they are offering an account of. Conservative constructivism seeks to conserve our judgements in the domain it offers an account of, while, of course, allowing room for some revisions. Radical constructivism does not have this objective. Rawls's constructivism about justice is an instance of conservative and bound constructivism. It employs normative notions, and seeks to preserve intuitive beliefs about justice.

Does Rawls have one constructivist project? No, there are two constructivist projects in Rawls. The first project adopts the point of view of a typical moral inquirer and asks what is the best theory of justice from her perspective. The second project seeks to identify a set of principles that will be acceptable to adherents of different reasonable comprehensive doctrines. The two projects are distinct because, as our discussion of the role of other people's beliefs showed, one can be rationally entitled to employ premises others disagree with in one's reasoning. Consequently, the contentious beliefs that are allowed in the first-person project are disallowed in the inter-personal project. In addition to this, beliefs that are disallowed in the first-person project may be allowed in the inter-personal project. For instance, one might on reflection think that whether a certain principle is unfeasible does not count against that principle being a correct principle of justice. Nevertheless, feasibility can play a legitimate role in the interpersonal project since its aim is identifying a set of principles to put into practice that will be acceptable to people who disagree on various issues.

Does Rawls reject Cohen's thesis that there are fact-insensitive principles? No, Rawls does not reject Cohen's thesis. Instead he holds that principles we are inclined to affirm without knowledge of facts may be true because certain facts obtain. In the terminology introduced in chapter 4, according to Rawls, metaphysically fact-insensitive principles and epistemically fact-insensitive principles do not always coincide. In fact, epistemically fact-insensitive principles may turn out to be metaphysically fact-sensitive. On Rawls's view, many of our beliefs about distributive justice, which can be epistemically fact-insensitive, are metaphysically fact-sensitive.

Does Rawls need to reject Cohen's thesis for his constructivist project? No, Rawls doesn't need to reject Cohen's thesis, because his project is an instance of bound and conservative constructivism. Because Rawls's theory is an instance of *bound* constructivism, he can acknowledge the existence of normative commitments in his constructivist device. And, because Rawls's constructivism is an instance of *conservative* constructivism, he can acknowledge the fact that the procedure is answerable to prior convictions about justice.

Can there be good reasons to characterize the decision problem in the original position as one under uncertainty without unjustifiably stipulating that the denizens of the original position lack probability estimates? Yes, there are good reasons for this characterization of the decision problem in the original position. These reasons are uncovered by examining the nature of the value function the denizens of the original position should hold, and our

cognitive capacities. The value function we should attribute to the denizens of the original position should violate the continuity axiom of decision theory. Continuity doesn't hold when one option is a life in which S is guaranteed the ability to do something with his life and the other option offers either (a) such a life plus some further goods, or (b) a life in which S cannot do something with his life, even if this latter possibility is very unlikely. However, the denizens of the original position cannot precisely identify the point where this discontinuity sets in. Accordingly, even if we stipulate that the denizens of the original position know the probabilities of ending up with any given amount of primary goods, they don't know the probability of ending up above or below the outcome where discontinuity sets in under different principles of distribution.

Is there a good case for an egalitarian ethos? When an egalitarian ethos is interpreted as a moral duty governing occupational choices that aims to bring about the co-realization of equality and Pareto-optimality, there is not a good case for it. Since the egalitarian ethos so construed is an instrument to bring about a certain end, we can evaluate it by asking whether the end is worth pursuing, and whether it can bring about this end. If primary goods are the metric of justice, in the case of primary goods that are governed by the difference principle, which are income and wealth, Pareto-optimality in the space of justice is not desirable. If actual preference satisfaction is the metric of justice, then the egalitarian ethos cannot bring about equality in this metric. If we are in a position to bring about equality in this space through other means, we will be in a position to bring about Pareto-optimality with-

out recourse to an egalitarian ethos. If the correct theory of welfare is the informed preference satisfaction view, or the objective list theory, and we rely on people's actual preferences as a proxy, then the previous point will apply. If we have measures in place that transforms people's actual preferences so that their actual preferences and what these theories recommend for them correspond, the egalitarian ethos will, again, be redundant. If we have an egalitarian ethos that's able to bring about Pareto-optimality in the space of welfare – when the correct theory of welfare is either the informed preference satisfaction theory or the objective list theory – even though people's actual preferences diverge from what these theories recommend for them, then this ethos will be objectionably paternalistic.

9.2 The Limits of this Defence and Some Qualifications

There are several limits and qualifications of the arguments I've presented. I would like to state them explicitly.

My defence in Part 3 only considered welfare and primary goods as possible candidates for the metric of justice. There are, however, other candidates. These metrics may not run into the same problems that the proposals for egalitarian ethos ran into when we assumed the metric of justice to be either welfare or primary goods. This is one way in which the argument of part 3 is limited. While arguing against the combination of a commitment to primary goods with W-Welfare even when welfare is understood to be ac-

tual preference satisfaction on page 197, I relied on Rawls's commitment to and his understanding of neutrality. This understanding of neutrality may be challenged. For instance, Arneson has argued that, contrary to Rawls's claim, a commitment to maximizing the satisfaction of informed and self-regarding preferences subject to distributive constraints is compatible with neutrality.[1] Arneson may be right about this. I have, nevertheless, not felt the need to address Arneson's challenge, because a large part of the appeal of primary goods, setting aside practical considerations, is a Rawlsian concern with neutrality. It would be odd to agree with Arneson on this point, and acknowledge the problem with pursuing Pareto-optimality in the space of primary goods, but not to switch to another metric. Another qualification of the argument of the third part is that it does not identify a problem with a moral duty that governs one's occupational choices as such. It identifies a problem either with the end for which the egalitarian ethos is a means, or with the effectiveness of the egalitarian ethos. Under different factual assumptions, an egalitarian ethos may very well be justified. More importantly, my argument does not apply to an egalitarian ethos more broadly construed. Certain attitudes and choices beyond complying with institutional demands are likely to be required by justice.

Some of the limitations of the argument of part 2 were already indicated there. I would like to emphasize again that the argument I've presented does not *clinch* the argument for maximin. Another limitation of the argument I've set out is the existence of problems with determining the value function of

[1]Richard J. Arneson, 'Primary Goods Reconsidered', *Nous*, 24 June (1990):3.

the denizens of the original position. As I pointed, the value function should not be specified in ways that makes the derivation of the principles trivial. In addition to this, we might want the value function to be acceptable to people with different conceptions of the good. The value function I attributed to the denizens of the original position maintained that there was a discontinuity between outcomes that afforded a decent life and the ones that failed to do this, though exactly which outcomes fell under which category was held to be uncertain. This is compatible with offering different accounts of what amounts to a decent life. Nevertheless, the discontinuity claim is a value claim, and may be open to challenge.

I would like to emphasize that my defence of Rawls in chapter 5 against Cohen's claim that Rawls's constructivism rests on a mistaken thesis about the relationship between facts and principles applies only to the constructivist project aimed to provide an account of justice that is best from the first-person perspective. It doesn't apply in the case of the constructivist project that has interpersonal justification as an aim. This can be seen by imagining the following case. Suppose according to a controversial economic theory, which I hold, and use in the derivation of the principles of justice in the original position, the principles that emerge out of the constructivist procedure will not give rise to inequalities that I find intuitively objectionable. In this case, I can apply the argument in chapter 5, and come to hold the theory derived in the way outlined in that chapter. I can come to judge that even though I held the belief that some inequalities were unjust without reference to any facts, their wrongness was metaphysically fact-sensitive. I

can't make the same move when the constructivism in question is of the kind that seeks interpersonal justification. Given that the economic theory that ensured this fit was controversial, it can't be used in this constructivist procedure. In this constructivist project, the denizens of the original position are allowed to hold only the *uncontroversial* claims of natural and social sciences.[2] I don't have a settled answer as to whether this is a problem for the kind of constructivism that has interpersonal justification as its aim. This is another limitation of the argument I've offered.

A related limitation of my argument is that I have not said anything about the significance of the differences between the principles generated by the kind of constructivism that seeks interpersonal justification and the one that seeks the theory that's best from the first-person perspective. The best case scenario is that the principles of justice according to the best theory of justice for a person and the principles that can be legitimately imposed coincide. The second best scenario would be one where the divergences are rare. However, if the recommendations of these two theories often conflict, then this person has to come up with reasons for why legitimacy should trump justice. This will require an account of the relationship between legitimacy and justice, and the over-riding significance of legitimacy. I have not offered this account.[3]

[2]John Rawls, *Political Liberalism*, New edition. (New York: Columbia University Press, 1996), pp. 67, 224-5; John Rawls, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999), p. 328-9.

[3]I don't mean to suggest that such an account cannot be given. For an idea how as to how we can build a response see Rawls, *Political Liberalism*, pp. 139-40, 218.

I have mentioned some of the qualifications of the argument I offered in chapter 3, but they are worth emphasizing again. I argued that we are not always required to respect Christensen's Independence principle. This, however, doesn't mean that we are always entitled to violate Independence. In the case of track record judgements from the ground level, for instance, if there are several instances of prior agreement, and the judgement in contention is not confidently held, then the more rational response would be lowering our confidence in the contentious belief. Since it's sometimes more rational to lower our confidence in our beliefs, the gap between what we can justify to others and what we are rationally entitled to believe may sometimes be narrow. If this is the case, the gap between the principles that can be legitimately imposed on others and the the best theory of justice for a person would also be narrower.

In chapter 2 I argued that metaethical theories that require revisions of our first-order beliefs face an uphill struggle, because we usually hold the beliefs that support metaethical theories with less confidence than our first-order beliefs. This does not mean that in the case of a conflict between a metaethical theory and a first-order belief, the first-order beliefs will *always* defeat the metaethical theory. It is possible that the first-order beliefs are not confidently held, in which case the metaethical theory can have enough weight to defeat the first-order beliefs. Furthermore, my claim applies to what I called typical moral inquirers. It is perfectly compatible with my account that some people place much more confidence in the beliefs that support metaethical theories than their first-order beliefs. For instance, philosophers

like Hare and Mackie may not fall under my characterization of typical moral inquirers in this respect.[4] The fact that we consider someone not to be a typical moral inquirer does not licence us to ignore the arguments for their views. Their arguments will employ premises we are inclined to accept, though perhaps not as confidently as them, and will reveal to us tensions in our views. The way we resolve these tensions may be different from how they would like us to resolve them, but their arguments will have a role in our reasoning.

9.3 Conclusion

The first part of this thesis examined the project of moral and political theorizing. One central lesson that has emerged from this examination is that there are several distinct projects in our theorizing, and these projects have different constraints. One difficulty that is unique to moral and political philosophy is that our thinking can have two guises. Moral philosophizing has a theoretical and a practical aspect. While we may disagree with others on

[4]Even though, I'm happy to grant that Hare and Mackie may not fall under my characterization of typical moral inquirers, I think there are reasons to think that even they are typical moral inquirers. In the case of Hare, even though he assigns a foundational role to his metaethical views, and chides Rawls for his reliance on our pre-theoretical first-order beliefs, he spends a lot of effort in showing that his theory matches our strongly held first-order beliefs. All of these aspect of Hare's thought are on display in R. M. Hare, 'Ethical Theory and Utilitarianism', in: Amartya Sen and Bernard Williams, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982). It is, of course, possible that his efforts are for the benefit of his interlocutors rather than himself. In the case of Mackie, even though he begins his *Ethics: Inventing Right and Wrong* with a presentation of his error-theory, which holds that our moral discourse presupposes the existence of moral properties in the world that do not actually exist, in the second half of the book he begins to present a moral theory that is not revisionary. In other words, as he sees it, his metaethical convictions don't commit him to a denial of his first-order views. J. L Mackie, *Ethics: Inventing Right and Wrong*, (London: Penguin, 1990).

what is the best moral theory, we may find that we agree on the course of action to take in specific instances. Given a commitment to the *pro tanto* badness of coercing people against their will, or a realization of the practical impossibility of putting principles into practice in the absence of the recognition of their legitimacy, much of our moral and political philosophizing will be aimed towards the practical goal of arriving at principles that we can all agree to follow. This will be a project that is intimately concerned with the beliefs of other people have. This is a project that has a lot of urgency. However, the theoretical project of arriving at the best account of justice, or some other question, is also a worthy endeavour. I have argued that Rawls has pursued both projects, but my defence has focused on Rawls's account of justice evaluated as a theoretical inquiry. I have suggested that its acceptance is less costly than it is claimed to be. His theory does not require us to adopt a specific account of justification, a metaethical view, or an egalitarian ethos, and one of its premises that appears *ad hoc* can be offered good grounds.

This is very far from establishing a 'moral geometry', and it is unlikely to change the minds of people who have illiberal views. Nevertheless, I hope it goes some way towards the wider project of defending Rawls's liberal egalitarian theory. It was unreasonable of me to expect political philosophy to accomplish the tasks of actual political struggles. Even so, having the vision of a just society in view can strengthen one's resolve.

Bibliography

Alchian, Armen A., 'The Meaning of Utility Measurement', *The American Economic Review*, 43 March (1953):1, pp. 26–50.

Alston, William P., 'Epistemic Circularity', *Philosophy and Phenomenological Research*, 47 September (1986):1, pp. 1–30.

Arneson, Richard, 'Welfare Should Be the Currency of Justice', *Canadian Journal of Philosophy*, 30 (2000):4, pp. 497–524.

Arneson, Richard J., 'Primary Goods Reconsidered', *Nous*, 24 June (1990):3, pp. 429–454.

Arrhenius, Gustaf, 'Superiority in Value', *Philosophical Studies*, 123 (2005):1, pp. 97–114.

Arrhenius, Gustaf and Rabinowicz, Wlodek, 'Value and Unacceptable Risk', *Economics and Philosophy*, 21 (2005):02, pp. 177–197.

Barry, Brian, *Theories of Justice: A Treatise on Social Justice, Vol. 1*, (Berkeley: University of California Press, 1989).

Bovens, Luc and Rabinowicz, Wlodek, 'Democratic Answers to Complex Questions - An Epistemic Perspective', *Synthese*, 150 May (2006):1, pp. 131–153.

Brandt, Richard B., *A theory of the Good and the Right*, Rev. ed. edition. (Amherst, N.Y.: Prometheus Books, 1998).

Broome, John, *Weighing Goods: Equality, Uncertainty and Time*, (Oxford: Basil Blackwell, 1991).

Buchanan, Allen, 'Justice, Legitimacy, and Human Rights', in: Davion, Victoria and Wolf, Clark, editors, *The Idea of a Political Liberalism: Essays on Rawls*, (Lanham: Rowman and Littlefield, 2000), pp. 73–89.

Campbell, Donald E., *Incentives: Motivation and the Economics of Information*, (Cambridge: Cambridge University Press, 1995).

Carens, Joseph H., *Equality, Moral Incentives, and the Market: An Essay in Utopian Politico-Economic Theory*, (Chicago: University of Chicago Press, 1981).

Carens, Joseph H., 'Rights and Duties in an Egalitarian Society', *Political Theory*, 14 (1986):1, pp. 31–49.

Carruthers, Peter, *Language, Thought and Consciousness: An Essay In Philosophical Psychology*, (Cambridge: Cambridge University Press, 1996).

Chipman, John S., 'The Foundations of Utility', *Econometrica*, 28 April (1960):2, pp. 193–224.

Chisholm, Roderick M., *Theory of Knowledge*, 2nd edition. (Englewood Cliffs: Prentice-Hall, 1977).

Christensen, David, 'Epistemology of Disagreement: The Good News', *Philosophical Review*, 116 (2007):2, pp. 187–217.

Christensen, David, *Disagreement, Question-Begging and Epistemic Self-Criticism*, Unpublished manuscript ⟨URL: http://www.brown.edu/Departments/Philosophy/onlinepapers/christensen/Conciliationism.pdf⟩.

Cleve, James van, 'Is Knowledge Easy - Or Impossible? Externalism as the Only Alternative to Skepticism', in: Luper, Steven, editor, *The Skeptics: Contemporary Essays*, (Aldershot: Ashgate, 2003), pp. 45–59.

Coady, C. A. J, *Testimony: A Philosophical Study*, (Oxford: Clarendon Press, 1992).

Cohen, G. A, *Karl Marx's Theory of History: A Defence*, Expanded edition. (Oxford: Oxford University Press, 2000).

Cohen, G. A, *If You're an Egalitarian, How Come You're so Rich?* (Cambridge, Mass: Harvard University Press, 2001).

Cohen, G. A, *Rescuing Justice and Equality*, (Cambridge, Mass: Harvard University Press, 2008).

Daniels, Norman, *Justice and Justification: Reflective Equilibrium in Theory*, (Cambridge: Cambridge University Press, 1996).

Dasgupta, Partha, 'Utilitarianism, Information and Rights', in: Sen, Amartya and Williams, Bernard, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982), pp. 199–218.

Dasgupta, Partha and Hammond, Peter, 'Fully Progressive Taxation', *Journal of Public Economics*, 13 April (1980):2, pp. 141–154.

Dorsey, Dale, 'Headaches, Lives and Value', *Utilitas*, 21 (2009):01, pp. 36–58.

Dworkin, Gerald, *The Theory and Practice of Autonomy*, (Cambridge: Cambridge University Press, 1988).

Dworkin, Ronald, 'Objectivity and Truth: You'd Better Believe it', *Philosophy and Public Affairs*, 25 (1996):2, pp. 87–139.

Elga, Adam, 'Reflection and Disagreement', *Nous*, 41 (2007):3, pp. 478–502.

Elga, Adam, *How to Disagree about How to Disagree*, Unpublished manuscript ⟨URL: http://philsci-archive.pitt.edu/archive/00003702/⟩.

Enoch, David, 'Can there be a global, interesting, coherent constructivism about practical reason?' *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, 12 (2009):3, pp. 319–339.

Enoch, David, *Not Just a Truthometer: Taking Oneself Seriously (but not too seriously) in Cases of Peer Disagreement*, Unpublished manuscript ⟨URL: http://law.huji.ac.il/upload/truthometer.pdf⟩.

Estlund, David, 'The Survival of Egalitarian Justice in John Rawls's Political Liberalism', *The Journal of Political Philosophy*, 4 (1996):1, pp. 68–78.

Estlund, David M., 'Opinion leaders, independence, and Condorcet's Jury Theorem', *Theory and Decision*, 36 March (1994):2, pp. 131–162.

Feldman, R. and Warfield, T., editors, *Disagreement*, (Oxford: Oxford University Press, Forthcoming).

Fine, Kit, 'Vagueness, Truth and Logic', in: Keefe, Rosanna and Smith, Peter, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997), pp. 119–150.

Fishburn, Peter C., *Nonlinear Preference and Utility Theory*, (Brighton: Wheatsheaf, 1988).

Fodor, Jerry A., *The Language of Thought*, (Cambridge, Mass: Harvard University Press, 1975).

Foley, Richard, *Intellectual Trust in Oneself and Others*, (Cambridge: Cambridge University Press, 2001).

Freeman, Samuel, 'Utilitarianism, Deontology, and the Priority of Right', *Philosophy and Public Affairs*, 23 (1994):4, pp. 313–349.

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

Griffin, James, 'Modern Utilitarianism', *Revue Internationale de Philosophie*, 141 (1982), pp. 331–375.

Griffin, James, *Well-Being: Its Meaning, Measurement and Moral Importance*, (Oxford: Clarendon Press, 1986).

Hare, R. M., 'What is Wrong with Slavery', *Philosophy and Public Affairs*, 8 (1979):2, pp. 103–121.

Hare, R. M., 'Ethical Theory and Utilitarianism', in: Sen, Amartya and Williams, Bernard, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982), pp. 23–38.

Harman, Gilbert, 'Three Trends in Moral and Political Philosophy', *The Journal of Value Inquiry*, 37 (2006):3, pp. 415–425.

Harsanyi, John C., 'Cardinal Utility in Welfare Economics and in the Theory of Risk-taking', *The Journal of Political Economy*, 61 (1953):5, pp. 434–435.

Hart, H. L. A, 'Rawls on Liberty and Its Priority', *University of Chicago Law Review*, 40 (1973):3, pp. 534–555.

Hausman, Daniel M. and McPherson, Michael S., *Economic Analysis, Moral Philosophy, and Public Policy*, 2nd edition. (New York: Cambridge University Press, 2006).

Hirose, Iwao, 'Review Article: Aggregation and Non-Utilitarian Moral Theories', *Journal of Moral Philosophy*, 4 July (2007):2, pp. 273–284.

Hooker, Brad, *Ideal Code, Real World: A Rule-Consequentialist Theory Of Morality*, (Oxford: Clarendon Press, 2000).

Hooker, Brad, 'Reflective Equilibrium and Rule Consequentialism', in: Hooker, Brad, Mason, Elinor and Miller, Dale E, editors, *Morality,*

Rules, and Consequences: A Critical Reader, (Edinburgh: Edinburgh University Press, 2000), pp. 222–238.

Howe, Roger E. and Roemer, John E., 'Rawlsian Justice as the Core of a Game', *The American Economic Review*, 71 December (1981):5, pp. 880–895.

Kamm, F. M., *Morality, Mortality: Death and Whom to Save From It*, Volume 1, (New York: Oxford University Press, 1993).

Katz, Jerrold J, *Realistic Rationalism*, (Cambridge, Mass: MIT Press, 1998).

Keefe, Rosanna, *Theories of Vagueness*, (Cambridge: Cambridge University Press, 2000).

Keefe, Rosanna and Smith, Peter, 'Introduction', in: Keefe, Rosanna and Smith, Peter, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997), pp. 1–57.

Kelly, Thomas, 'The Epistemic Significance of Disagreement', in: Hawthorne, John and Szabo, Tamar Gendler, editors, *Oxford Studies in Epistemology*, Volume 1, (Oxford: Clarendon Press, 2005), pp. 167–196.

Kelly, Thomas, *Peer Disagreement and Higher Order Evidence*, Unpublished manuscript ⟨URL: http://www.princeton.edu/~tkelly/papers/KellyPeerDis-1.doc⟩.

Kornblith, Hilary, *Epistemology: Internalism and Externalism*, (Malden, Mass: Blackwell Publishers, 2001), p. 271.

Korsgaard, Christine M, *The Sources of Normativity*, (Cambridge: Cambridge University Press, 1996).

Kreps, David M., *Notes on the Theory of Choice*, (Boulder, Colo: Westview, 1988).

Kymlicka, Will, 'Rawls on Teleology and Deontology', *Philosophy and Public Affairs*, 17 (1988):3, pp. 173–190.

Lackey, Jennifer, *A Justificationist View of Disagreement's Epistemic Significance*, Unpublished manuscript ⟨URL: http://faculty.wcas.northwestern.edu/~jal788/documents/JustificationistDisagreement-SE.pdf⟩.

Laffont, Jean-Jacques and Martimort, David, *The Theory of Incentives: The Principal-Agent Model*, (Princeton, N.J: Princeton University Press, 2002).

Larmore, Charles E., *Patterns of Moral Complexity*, (Cambridge: Cambridge University Press, 1987).

Lehrer, Keith, *Theory of Knowledge*, 2nd edition. (Boulder, Colo: Westview Press, 2000).

Lewis, David K., *On the Plurality of Worlds*, (Oxford: Blackwell, 2001).

List, Christian, 'The Discursive Dilemma and Public Reason', *Ethics*, 116 (2006):2, pp. 362–402.

Luce, R. Duncan and Raiffa, Howard, *Games and Decisions: Introduction and Critical Survey*, (New York: Dover, 1989).

Lycan, William G., 'Moore Against The New Skeptics', *Philosophical Studies*, 103 (2001):1, pp. 35–53.

Mackie, J. L, *Ethics: Inventing Right and Wrong*, (London: Penguin, 1990), p. 249.

Mill, John Stuart, 'Utilitarianism', in: *Collected Works*, Volume 10, (Toronto: University of Toronto Press, 1991).

Miller, David, *Social Justice*, (Oxford: Clarendon Press, 1976), p. 367.

Mirrlees, J. A., 'An Exploration in the Theory of Optimum Income Taxation', *The Review of Economic Studies*, 38 (1971):2, pp. 175–208.

Nagel, Thomas, 'Rawls on Justice', *The Philosophical Review*, 82 April (1973):2, pp. 220–234.

Nagel, Thomas, *The Last Word*, (New York: Oxford University Press, 1997).

Nagel, Thomas, 'Equality', in: Williams, Andrew and Clayton, Matthey, editors, *The Ideal of Equality*, (Basingstoke: Palgrave MacMillan, 2002), pp. 60–80.

Neumann, John Von and Morgenstern, Oskar, *Theory of Games and Economic Behavior*, 3rd edition. (Princeton, N.J.: Princeton University Press, 1953).

Norcross, Alastair, 'Comparing Harms: Headaches and Human Lives', *Philosophy and Public Affairs*, 26 (1997):2, pp. 135–167.

Nozick, Robert, *Philosophical Explanations*, (Oxford: Clarendon Press, 1981), p. 764.

Nozick, Robert, *The Nature of Rationality*, (Princeton: Princeton University Press, 1993).

Olsson, Erik J, *Against Coherence: Truth, Probability, and Justification*, (Oxford: Oxford University Press, 2005), p. 232.

O'Neill, Onora, *Constructions of Reason: Explorations of Kant's Practical Philosophy*, (Cambridge: Cambridge University Press, 1989).

Otsuka, Michael, 'Saving Lives, Moral Theory, and the Claims of Individuals', *Philosophy & Public Affairs*, 34 (2006):2, pp. 109–135.

Parfit, Derek, *On What Matters*, ⟨URL: http://users.ox.ac.uk/~ball2568/parfit/parfit_-_on_what_matters.pdf⟩.

Parfit, Derek, 'Overpopulation and the Quality of Life', in: Singer, Peter, editor, *Applied Ethics*, (Oxford: Oxford University Press, 1986), pp. 145–164.

Parfit, Derek, 'Equality and Priority', *Ratio*, 10 (1997):3, pp. 202–221.

Pettit, Philip, 'Deliberative Democracy and the Discursive Dilemma', *Philosophical Issues (supplement to Nous)*, 11 (2001), pp. 268–295.

Pettit, Philip, 'When to defer to majority testimony - and when not', *Analysis*, 66 July (2006):3, pp. 179–187.

Pettit, Philip and Rabinowicz, Wlodek, 'The Jury Theorem and the Discursive Dilemma.' *Philosophical Issues (supplement to Nous)*, 35 (2001), pp. 295–299.

Plato; Cooper, John M. and Hutchinson, D. S, editors, *Plato: Complete Works*, (Indianapolis, IN: Hackett Publishing Company, 1997).

Pollock, John, 'Epistemic Norms', in: Sosa, Ernest and Kim, Jaegwon, editors, *Epistemology: An Anthology*, (Malden, Mass: Blackwell, 2000), pp. 192–225.

Qizilbash, Mozaffar, 'Transitivity and Vagueness', *Economics and Philosophy*, 21 (2005):01, pp. 109–131.

Rabinowicz, Wlodek, 'Discussion - Ryberg's Doubts about Higher and Lower Pleasures - Put to Rest?' *Ethical Theory and Moral Practice*, 6 (2003):2, pp. 231–235.

Railton, Peter, 'Moral Realism', *The Philosophical Review*, 95 (1986):2, pp. 163–207.

Rawls, John, *Political Liberalism*, New edition. (New York: Columbia University Press, 1996).

Rawls, John, *Collected Papers*, (Cambridge, Mass: Harvard University Press, 1999).

Rawls, John, *A Theory of Justice*, Rev. edition. (Oxford: Oxford University Press, 1999).

Rawls, John, *Lectures on the History of Moral Philosophy*, (Cambridge, Mass: Harvard University Press, 2000).

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

Russell, Bertrand, 'Vagueness', in: Keefe, Rosanna and Smith, Peter, editors, *Vagueness: A Reader*, (Cambridge, Mass: MIT, 1997), pp. 61–68.

Scanlon, T. M., 'Contractualism and Utilitarianism', in: Sen, Amartya and Williams, Bernard, editors, *Utilitarianism and Beyond*, (Cambridge: Cambridge University Press, 1982), pp. 199–218.

Scanlon, Thomas, *What We Owe to Each Other*, (Cambridge, Mass: Belknap Press of Harvard, 1998).

Schmidt, Ulrich, *Axiomatic utility theory under risk: Non-Archimedean Representations and Application to Insurance Economics*, (Berlin: Springer, 1998).

Sen, Amartya, *Inequality Reexamined*, (New York: Russell Sage Foundation, 1992).

Sen, Amartya, 'Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms', *Oxford Economic Papers*, 45 October (1993):4, pp. 519–541.

Sen, Amartya and Foster, James E., *On Economic Inequality*, Enl. edition. (Oxford: Clarendon Press, 1997).

Sidgwick, Henry, *The Methods of Ethics*, 2nd edition. (London: Macmillan, 1877).

Sidgwick, Henry, *Essays on Ethics and Method*, (Oxford: Clarendon Press, 2000).

Soames, Scott, *Philosophical Analysis in the Twentieth Century*, Volume 1, (Princeton, N.J: Princeton University Press, 2003).

Sorensen, Roy A., 'Vagueness within the Language of Thought', *The Philosophical Quarterly*, 41 October (1991):165, pp. 389–413.

Stich, Stephen, 'Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity', *Synthese*, 74 March (1988):3, pp. 391–413.

Street, Sharon, 'Constructivism about Reasons', in: Shafer-Landau, Russ, editor, *Oxford Studies in Metaethics*, Volume 3, (Oxford: Oxford University Press, 2008), pp. 207–246.

Temkin, Larry, 'Worries about Continuity, Transitivity, Expected Utility Theory, and Practical Reasoning', in: Egonsson, Dan et al., editors, *Exploring Practical Philosophy*, (Aldershot: Ashgate, 2001), pp. 95–108.

Tuck, Richard, 'Is There a Free Rider Problem?' in: Harrison, Ross, editor, *Rational Action: Studies in Philosophy and Social Science*, (New York: Cambridge University Press, 1979), pp. 147–156..

Tuomala, Matti, *Optimal Income Tax and Redistribution*, (Oxford: Clarendon Press, 1990).

Waldron, Jeremy, 'Theoretical Foundations of Liberalism', *The Philosophical Quarterly*, 37 April (1987):147, pp. 127–150.

Wedgwood, Ralph, *The Nature of Normativity*, (Oxford: Clarendon Press, 2007).

Wilkinson, T. M., *Freedom, Efficiency and Equality*, (Basingstoke: Macmillan, 2000).

Williams, Andrew, 'Incentives, Inequality, and Publicity', *Philosophy and Public Affairs*, 27 (1998):3, pp. 225–247.

Williams, Bernard A. O., 'Internal and External Reasons', in: *Moral Luck: Philosophical Papers, 1973-1980*, (Cambridge: Cambridge University Press, 1981), pp. 101–113.

Williamson, Timothy, *Vagueness*, (London: Routledge, 1994).