Reconstructing Rawls: A Utilitarian Critique of Rawls's Theory of Justice

By

Samuel Patrick Fremantle

Submitted for the Phd in Philosophy at UCL

I, Samuel Patrick Fremantle, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

My thesis argues that Rawls's attempt to discredit utilitarianism as a viable theory of justice was ultimately unsuccessful. I shall follow the example of Robert Paul Wolff's 1977 book *Understanding Rawls* in treating *A Theory of Justice* 'not as a single piece of philosophical argument to be tested and accepted or rejected whole, but as a complex, many-layered record of at least twenty years of philosophical growth and development'. Paying close attention to the wording of different variants of Rawls's arguments as they developed over the years, I shall reconstruct my own argument using the most coherent parts of Rawls's arguments, along with contributions from various commentators. This will uphold the classical principle of utility, as a principle of distributive justice that is *entirely suited* to Rawls's conception of society as a cooperative venture for mutual advantage, with that conception's commitment to conceiving obligations of justice as essentially obligations of reciprocity. In doing so, I hope to show that the case against utilitarianism is unproven as is the case that justice requires the recognition of inviolable rights. My argument should also explain Rawls's continued modification of his arguments as largely due to his failure to successfully refute utilitarianism.

Acknowledgements

It is customary to acknowledge the support of one's supervisor, but I think in my case it's fair to say that I owe an extraordinary debt of gratitude to my long-suffering supervisor Professor Mike Otsuka, not just for his excellent criticisms and comments on my thesis (as many footnotes testify) but for his personal support for me which went well beyond the call of duty – on more than one occasion. I think it's highly unlikely that I would have managed to submit a thesis without him, and I count myself very lucky indeed to have had his help when I needed it.

I had an interesting and productive viva with my examiners, Dr Avia Pasternak and Dr Alex Voorhoeve, who produced a very thorough and considered examiners' report. I believe my resubmitted thesis is much improved as a result of their comments.

I also owe a huge debt of gratitude to my parents who supported me morally and financially when I was having a difficult time.

Other people I would like to mention for their support include Ben Short, Mark Fielding, Dr Cathy Greenwood, James Wilson, Raj Sehgal and Dan Adams.

Last, but not least, I'd like to acknowledge my sister, Joanna Fremantle (1964 – 1986), who wanted me to succeed and prosper, and whose life, relationship with me, and tragic death had such a profound effect on my ethical stance, especially my interest in, and position on, justice.

Table of Contents

Though society is not founded on a contract, and though no good purpose is answered by inventing a contract in order to deduce social obligations from it, every one who receives the protection of society owes a return for the benefit, and the fact of living in society renders it indispensable that each should be bound to observe a certain line of conduct towards the rest.

- John Stuart Mill, *On Liberty*, 1859.

Preface

In John Rawls's Preface to the original edition of *A Theory of Justice* (1971), Rawls wrote

Passage P1 ***(T of J orig)***

> During much of modern moral philosophy the predominant systematic theory has been some form of utilitarianism…Those who criticized them [i.e. the great utilitarians such as Hume, Smith and Mill] often did so on a much narrower front. They pointed out the obscurities of the principle of utility and noted the apparent incongruities between many of its implications and our moral sentiments. But they failed, I believe, to construct a workable and systematic moral conception to oppose it. The outcome is that we often seem forced to choose between utilitarianism and intuitionism. Most likely we finally settle upon a variant of the utility principle circumscribed and restricted in certain ad hoc ways by intuitionistic constraints. Such a view is not irrational; and there is no assurance that we can do better. But this is no reason not to try.
>
> What I have attempted to do is to generalize and carry to a higher order of abstraction the traditional theory of the social contract as represented by Locke, Rousseau, and Kant. In this way I hope that the theory seems to offer an alternative systematic account of justice that is superior, or so I argue, to the dominant utilitarianism of the tradition. The theory that results is highly Kantian in nature. - *A Theory of Justice* (1971)[1]

Although Rawls seldom explicitly described it as such until later works, his theory is, in my opinion, best thought of as a theory of 'justice as reciprocity'.[2] Justice as Reciprocity

[1] Rawls 1971 p. viii

[2] Rawls 1995 p.17. Rawls first described his theory of Justice as Fairness as one of Justice as Reciprocity in the first of his essays entitled 'Justice as Fairness' (1957 p. 661) In *Political Liberalism* Rawls endorsed Allan Gibbard's reading of his theory as such, in preference to Brian Barry's interpretation of it as 'hovering uneasily' between 'justice as impartiality' and 'justice as

construes our obligation to act 'justly' as one of constraining our behaviour in accordance with certain rules of conduct in order to give fair return to others for the benefits we receive from the similar constraint of others. The point behind Rawls's generalizing and carrying 'to a higher order of abstraction the traditional social contract theory' was to work out what the rules conforming to the conception of Justice as Reciprocity should be. Rawls named his idea that the correct rules to govern society conceived of as a cooperative venture for mutual advantage were those that would be chosen in the appropriately constructed social contract, 'justice as fairness'.

Utilitarianism, by contrast, Rawls conceived of as stemming from the conception of 'Justice as Benevolence'. He never described it explicitly in those actual words,[1] but that description seems to me to fairly capture his view of the matter. Rawls's conception of 'Justice as Benevolence' is captured in the passages repeated below. The first is taken from his book *Justice as Fairness: A Restatement* (2001), which, as the title suggests, was a restatement of the contract view of *A Theory of Justice,* albeit with some significant revisions, which was published shortly before Rawls's death. The second is from 'Justice as Fairness' (1958)

Passage P2 ***(J as F:AR 2001)***

> In the history of democratic thought two contrasting ideas of society have a prominent place: one is the idea of society as a fair system of social cooperation between citizens regarded as free and equal; the other is the idea of society as a social system organized so as to produce the most good summed over all its members, where this good is a complete good specified by a comprehensive doctrine. The tradition of the social contract elaborates the first idea, the utilitarian tradition is a special case of the second.
>
> Between these two traditions there is a basic contrast: the idea of society as a fair system of social cooperation is quite naturally specified so as to include the ideas of equality (the equality of basic rights, liberties, and fair opportunities) and of reciprocity (of which the difference principle is an example). By contrast, the idea of society organized to produce the most good expresses a maximizing and aggregative principle of political justice. In

---

mutual advantage'. (Rawls 1995 p. 17) He repeated this endorsement in *Justice as Fairness: A Restatement.* (Rawls 2001 p. 49). See also Barry 1989 and Gibbard 1991.

[1] I have not come across this terminology elsewhere in the literature, so as far as I know this name for a 'conception of justice' originated here.

> utilitarianism, the ideas of equality and of reciprocity are accounted for only indirectly, as what is thought to be normally necessary to maximize the sum of social welfare. - *Justice as Fairness: A Restatement* (2001)[1]

Passage P3 ***(J as F 1957)***

> This conception of justice (i.e. the conception of justice as reciprocity) differs from that of the stricter form of utilitarianism (Bentham and Sidgwick), and its counterpart in welfare economics, which assimilates justice to benevolence and the latter in turn to the most efficient design of institutions to promote the general welfare.[2]

These three passages should provide enough material for me to set out the main thrust of my argument in this thesis. 'Justice as Benevolence' would view the obligations of justice as a matter of our being obliged to produce as much good as possible. Classical utilitarianism is, to use the terms of **Passage P2**, 'a special case' of the conception of 'Justice as Benevolence' which defines the good in terms of happiness, and thus holds that we are obliged to produce as much happiness as possible. And, as Rawls observes in various places, in answer to the question of how we should be motivated to produce as much happiness as possible, utilitarians tend to appeal to the power of sympathy.[3] We feel pain at the prospect of others feeling pain, and pleasure at the prospect of others feeling pleasure, and these feelings impel us to promote the greatest aggregate of pleasure over pain overall.[4] But, Rawls maintained, while that may be a laudable way to behave that is not what justice is about. Justice is about being motivated to return benefits fairly to the people who you owe such return to.

It should be clear from the forgoing, I hope, that 'Justice as Benevolence' and 'Justice as Reciprocity' draw on two very different motives, and so, in advance of an argument to the contrary, might be expected to prescribe very different principles of distribution for society. The overriding aim of this thesis is to provide that argument to the contrary, and I shall do so by attempting to argue that Justice as Reciprocity is fully reconcilable with adoption of the principle of utility, interpreted on classical utilitarian

[1] Rawls 2001 p.96
[2] Rawls 1957 p.660
[3] Rawls 1999 pp. 24, 155, 162, 426
[4] E.g. see Mill 2003 p. 204

lines, as the predominant principle of distribution for society.[1] Two analogies will help to clarify this aim, and my forthcoming argument, here. Imagine someone who had never studied economics before being told that individuals who were purely motivated by the pursuit of their self-interest would actually behave in the way most conducive to promoting the public good. Their initial reaction might be one of incredulity. But Adam Smith in *The Wealth of Nations* demonstrated that individual pursuit of self-interest could be the most effective way of promoting the good of all, given the right market conditions and observance of the rule of law acting as a constraint on self-interested behaviour. As he put it, 'It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest.'[2]

Adam Smith's point was that the butcher, brewer and baker do not act from the predominant motive of providing the public with good quality, and affordable, meat, beer and bread. Instead their predominant motive is to maximize their own profit. But seeking to promote their own interest leads them to compete with each other in terms of quality and price, and so they end up promoting the good of society at large, though that was not the main motivation behind their behaviour.

The first analogy with my argument is this. Imagine an 'ideal legislator' charged with the task of designing institutions, which might include both legal restraints and customs, most suited to the conception of Justice as Benevolence, that is, most suited to promoting the greatest overall good.[3] Just as it might seem unlikely to the person ignorant of economics that people motivated by self-interest might nonetheless effectively promote the general good, so it might seem to a political philosopher approaching the issue in question, highly unlikely that an ideal legislator whose motivation was to come up with rules suited for the conception of Justice as Benevolence might nonetheless produce rules suited for the conception of Justice as Reciprocity. I think it is fair to speculate that Rawls approached the issue, at least at first, with a strong conviction that the two conceptions of

[1] I have been careful to specify a 'classical interpretation of the principle of utility' rather than 'classical utilitarianism' here, as classical utilitarianism, as ordinarily understood, is not reconcilable with justice as reciprocity without an important modification, as I shall explain later.
[2] Smith 2007 p.10
[3] The ideal legislator is a figure evoked by Rawls in his essays 'Justice as Fairness' and in *A Theory of Justice*, to explain how utilitarianism, as he interprets it, might be derived.

justice would inevitably yield drastically different policy recommendations.[1] But, I shall argue, they don't.

> A similar analogy concerns the behaviour of citizens in the just society. The motivation of citizens, according to Justice as Reciprocity, should not be to act out of sympathy and promote the general good, but to return benefits fairly to society in return for the good that society has done for them. But, I shall argue, if people were to ask themselves what principle an ideally benevolent person (who is motivated by sympathy, rather than reciprocity) would endorse, and then to seek to have that principle adopted as the predominant principle of distribution for society, the chosen principle would still be entirely compatible with the conception of Justice as Reciprocity.

I won't succeed in providing a watertight case that Justice as Reciprocity mandates the acceptance of the classical principle of utility, and the second to last sentence of the paragraph above can be used to provide some indication as to why not. There is nothing inherent in the notion of an 'ideally benevolent person' which commits such a person to classical utilitarianism. It could be, and has been, argued that ideal benevolence commits us to prioritizing the worst off.[2] And, as already remarked, within those theories that do stipulate that benevolence should require maximization of the good, utilitarianism is only a 'special case'.[3] It is beyond the scope of this thesis to fully consider the question of whether the ideally benevolent person should choose to maximize happiness, or do something else. Instead I shall argue that *if* the ideally benevolent person would choose to be a classical utilitarian *then* the classical principle of utility (as I shall call it from now on) is reconcilable with Justice as Reciprocity. This argument is enough for my central purpose of demonstrating that the classical principle of utility is compatible with the conception of justice as reciprocity. I may as well state up front what will become quite apparent with the progress of this thesis, which is that I am quite sympathetic towards the classical principle of utility myself. It has, it seems to me, major advantages in dealing with the problem of

---

[1] Though he does, at times, admit to nagging doubts that they won't. E.g. Rawls *T of J Rev* 1999 p.24.
[2] Notably by Thomas Nagel in *The Possibility of Altruism.* Although Nagel writes of 'altruism' rather than 'benevolence', the difference in terminology appears to me to be insignificant.
[3] John Broome has argued in *Weighing Goods* that 'teleology' with its commitment to maximization can accommodate a wide variety of precepts such as fairness, that are not accommodated by classical utilitarianism.

intergenerational justice, the severely disabled, and non-human animals, all of which are poorly dealt with by Rawls's theory of justice. It has also, I think, been treated very unfairly recently at the hands of its critics, particularly, as I shall demonstrate, John Rawls. I hope to redress the balance to some extent.

Championing the classical principle of utility in this way should make some important points even to those who remain entirely unsympathetic to the principle. Showing the classical principle of utility to be reconcilable with Justice as Reciprocity, should show that Rawls failed to establish his important, and influential, claim; that to take the notion of justice seriously involves the acknowledgment of inviolable rights.

In fact, I shall argue that, despite Rawls's failure to establish that claim, taking the notion of justice seriously *does* involve the recognition of inviolable rights, though not to nearly such an extensive degree as most believers in rights would like. This recognition is forced by the need to reconcile the classical principle of utility with the demands of Justice as Reciprocity. Classical act-utilitarianism places all agents, no matter what position they might be in, under an obligation to perform one of those acts that would maximize utility. I shall argue in Chapter 4 that the logic of Justice as Reciprocity grants those who would benefit insufficiently from their cooperation in a society ordered by the classical principle of utility the right to not cooperate. This significant modification of traditional classical act-utilitarianism is sufficient to justify my labelling the position arrived at, 'Reciprocal Classical Utilitarianism.'

My approach in this thesis has been heavily influenced by Robert Paul Wolff's book *Understanding Rawls* (1977).[1] I had been struggling, and failing, to make sense of some key passages where Rawls argued against utilitarianism. Wolff read *A Theory of Justice* 'not as a single piece of philosophical argument to be tested and accepted or rejected whole, but as a complex, many-layered record of at least twenty years of philosophical growth and development', continuing that '[t]he labyrinthine complexities of *A Theory of Justice* are the consequences of at least three stages in the development of Rawls's thought, in each of which he complicated his theory to meet objections others had raised to earlier versions, or which he himself perceived.'[2] Wolff also remarks on the 'numerous serious inconsistencies

[1] Jonathan Wolff first pointed out to me that my approach to A Theory of Justice was similar to Robert Paul Wolff's and that I would profit by reading it.
[2] Wolff 1977 p.4

and unclarities that make it appear that Rawls could not make his mind up on some quite fundamental questions'.[1] This reading shed considerable light on the difficulties I had faced in understanding *Theory*. I couldn't make sense of certain passages because they *didn't make sense*. They were the result of attempts to patch up arguments he had put forward in previous essays, by substituting new premises for old ones. It was possible to see this by comparing passages from *Theory* side by side with the ones that were their obvious antecedents from his earlier essays. But the problem was that the new premises didn't provide effective patches.

My diagnosis of the fundamental problem with Rawls's theory, however, is different to Wolff's.[2] As might be predictable on the basis of the preceding paragraphs, my explanation for Rawls's continuous revision of his arguments - which went on for another three decades after the publication of *Theory (1971),* to the publication of *Justice as Fairness: A Restatement (2001),* is that he was never able to defeat classical utilitarianism. On the basis of his first model (as Wolff refers to it[3]), it must have looked as though the claims Rawls was obviously determined to make in favour of his principles of justice and against the principle of utility could be easily upheld. The subsequent revisions of his theory should be read as (sometimes quite desperate) attempts to invent new ways to make his charges against the classical principle of utility stick.

In order to make good this case it is necessary for me to devote much more attention to the actual wording of Rawls's arguments than might normally be regarded as interesting or productive with so modern a philosopher. But the specific wording of Rawls's arguments is revelatory of the problems that he was struggling with, and by analysing the wording of earlier arguments and comparing it with that of later versions it is possible to discern the underlying reasons motivating his explicit stance.[4]

---

[1] Wolff 1977 p.3

[2] Wolff argued that Rawls's reliance on 'formal models of analysis drawn from the theory of rational choice' and 'the use of the concepts and models of utility theory, welfare economics and game theory' was 'fundamentally wrong' and 'the wrong way to deal with the normative and explanatory problems of social theory' (Wolff 1977 p.10)

[3] Wolff 1977 p.25 Wolff identifies and describes three models of Rawls's theory in *Understanding Rawls*. I believe I have identified another model in Rawls's essay 'Constitutional Liberty and the Concept of Justice' (1963) which lies halfway between Wolff's first and second models. Appropriately enough, I refer to it as model 1.5 and its distinguishing features play an important role in my argument of Chapter 4.

[4] This reading of Rawls again echoes Robert Paul Wolff's reading. Wolff remarks of Rawls's

There is a secondary purpose to my preoccupation with the wording of Rawls's arguments. If, as I hope to show, his 'official' arguments against utilitarianism are unsuccessful, the influence his work has undoubtedly had in persuading people that utilitarianism is incompatible with distributive justice, is largely due to the rhetorical and emotive language his arguments are couched in. By revealing it to be such, I hope to undo some of the damage.

Because of my close attention to pointing out various inconsistencies and incoherencies in the wording of Rawls's works, I did for a time toy with the title 'Unravelling Rawls' for my thesis. But as I also have the more positive aim of demonstrating that Rawls's conception of Justice as Reciprocity can be reconciled with, what I believe to be an attractive principle of benevolence, the classical principle of utility, I settled for a correspondingly positive title for my thesis: *Reconstructing Rawls: A Utilitarian Critique of Rawls's Theory of Justice.*

---

engagement in 'elaborate speculative moral psychology' in justification of his stipulation that the parties in the original position would not be prone to envy, that 'those speculations are strictly *post hoc.* The real reason for the assumption of non-envy is purely technical, and has to do with the assumptions required by the modes of quasi-economic reasoning that Rawls wishes to deploy.'(Wolff 1977 p. 29)

Chapter by Chapter Outline of Thesis

Chapter 1, entitled 'The Promise' can be thought of as an introductory chapter. explains the important, and complicated, concepts that the rest of the thesis will deal with, including the core conceptions of justice, Justice as Benevolence and Justice as Reciprocity; Rawls's conception of Justice as Fairness; Rawls's two principles of justice and classical utilitarianism and the 'three tenets of deontological liberalism', as I shall refer to them.[1]

The chapter also explains the very real attraction of Rawls's project in terms of its 'promise' to account for the tenets of deontological liberalism by grounding them firmly in the conception of Justice as Reciprocity.

But Chapter 2 will go on to expose problems for Rawls's project that can be discerned by inconsistencies and unclarities in the text of *A Theory of Justice*; problems that I claim Rawls was never able to resolve satisfactorily.

In Chapter 3, I examine closely what Robert Paul Wolff called the 'first form' of Rawls's model, as set out in Rawls's two essays entitled 'Justice as Fairness' (1957 & 1958). The purpose of this examination will be to demonstrate how much more effective the first form of Rawls's model would have been than subsequent models in sustaining two separate, but related, claims. First, that the conception of Justice as Reciprocity would only be compatible with Rawls's two principles of justice and secondly; that the classical principle of utility would likely *not* be compatible with the conception of Justice as Reciprocity – it would, in fact, only be compatible with the conception of Justice as Reciprocity in the circumstance that the policy recommendations of the classical principle of utility exactly coincided with the policy recommendations of the two principles of justice.[2]

[1] The three tenets of deontological liberalism will be fully explained in Chapter 1.
[2] My assertions here that the classical principle of utility would probably not be compatible with the conception of Justice as Reciprocity and would be unlikely to coincide with the policy recommendations of the two principles of justice is intended to capture the spirit of Rawls's arguments in his essays of the first model. I suggest below (§1.5.2) that it is quite likely that the classical principle of utility would yield similar policy prescriptions to the two principles of justice.

So Chapter 3 will have left Justice as Reciprocity in need of a tiebreaker between the two principles of justice and the principle of classical utility. Chapter 3 will undertake an investigation of Rawls's repeated, and varied, claims in his essays and books – starting with 'Distributive Justice' (1967) and continuing through the original and revised editions of *A Theory of Justice* (1971& 1999) to *Justice as Fairness: Restatement* (2001) – that the principle of utility would clash with the conception of Justice as Reciprocity by violating some *other* requirement of reciprocity than Rawls's central claim that it would not be chosen by the parties in a hypothetical contract. I shall argue that these other requirements suggested by Rawls are not plausible requirements of the conception of Justice as Reciprocity.

In Chapter 3 I shall also conclude the argument, started in Chapter 1 and continued in Chapter 2, that 'Justice as Fairness' – Rawls's name for his hypothetical contract device – is unfit for the purpose of constructing principles suited to the conception of Justice as Reciprocity.

The position my argument will have reached by the end of Chapter 3 will be that Rawls failed to demonstrate that the classical principle of utility is incompatible with the conception of Justice as Reciprocity. Relatedly, he also failed to demonstrate that the demands of Justice as Reciprocity diverge from the demands of Justice as Benevolence.

In Chapter 4 I argue that the classical principle of utility is reconcilable with Justice as Reciprocity. For this purpose, I do not need to show that the classical principle of utility is *uniquely* suited to Justice as Reciprocity. I just need to maintain that it is, as Rawls concedes it is, *a* reasonable conception of the good.[1] This is enough, I shall argue, to allow an 'ideal legislator', charged with the task of choosing principles suited to the conception of Justice as Reciprocity to select the classical principle of utility, without fear that she or he

[1] Rawls acknowledges the utilitarianism of Bentham and Sidgwick as a 'reasonable comprehensive doctrine' in 'The Idea of an Overlapping Consensus.' (Rawls 1996 pp. 169-170)

would violate any requirement of Justice as Reciprocity.[1]

However, I shall go on to suggest that the selection of the classical principle of utility by the Ideal Legislator (Reciprocity) does not result in classical utilitarianism as it is ordinarily understood. This is because classical utilitarianism, as it is ordinarily understood, obliges everyone, no matter what position they might be in, to always act so as to maximize the greatest happiness of the greatest number. I shall argue that Justice as Reciprocity, when combined with the selection of the classical principle of utility, would exempt some people from that obligation. It is beyond the remit of this thesis to decide the question of who precisely should be exempted from this obligation, though I will put forward some considerations on this point. But for the purpose of the argument of this thesis, I do not need to resolve this question. This is because whoever those who should be exempted from the obligation to maximize utility might turn out to be, their exemption will not interfere with the right of others to maximize utility. And to show that people would have this right is enough to show that the classical principle of utility is reconcilable with Justice as Reciprocity.[2]

The justification for the exemption of some from the obligation to maximize utility is based in the conception of Justice as Reciprocity. Justice as Reciprocity provides three potential reasons for exempting certain people from this obligation. The first is that some people may be psychologically incapable of cooperating, in which case they do not qualify as cooperating members of society to whom the obligation of reciprocity applies. The second is that they may not be in the circumstances of justice. And the third is that the 'benefits' that would accrue to them from their cooperation may not be sufficient for their cooperation to count as 'advantageous' to them. All three of these reasons for exemption point, I

[1] The 'ideal legislator (Reciprocity)' is an abbreviation of 'the ideal legislator charged with the task of choosing principles suited to the conception of Justice as Reciprocity. This is a character I shall introduce in Chapter 4 for the purpose of explaining the argument of that chapter. It is based on the 'ideal legislator' referred to by Rawls in *Theory*, whose task it is to design institutions suited to the conception of Justice as Benevolence.

[2] The reason that the exempted people's right to not cooperate does not interfere with others' rights to maximize utility is because their right is a 'liberty-right' not a 'claim-right'. The distinction between these two types of right is explained below (§1.1.2). This summary of the argument of Chapter 4 should be more comprehensible after the explanation of this distinction and the argument of Chapter 4.

believe, to a ‘threshold’ view, where the obligation to fully cooperate kicks in after cooperation affords a certain level of overall life prospects.

The resulting conception of justice I call ‘Reciprocal Classical Utilitarianism’ and differs in its demands from classical utilitarianism only in allowing for the exemption just described. But this difference is not as insignificant as it might first appear. The argument that the Ideal Legislator (Reciprocity) has the right to select reciprocal classical utilitarianism but not classical act-utilitarianism will demonstrate that Rawls was right on two important points. Firstly, that Rawls turned out to be right in his conviction that the demands of Justice as Benevolence would differ from the demands of Justice as Reciprocity. Secondly, that the conception of Justice as Reciprocity does uphold the core tenet of deontological liberalism; it would grant some people inviolable rights that classical utilitarianism would be prone to violate.

But this result is a disappointing one as judged from the perspective of the initial promise of Rawls’s project. That, as I demonstrate in Chapter 1, would have upheld all three tenets of deontological liberalism and decisively rejected the classical principle of utility. Instead, I shall conclude, Justice as Reciprocity is reconcilable with the classical principle of utility and the remaining two tenets of deontological liberalism remain unaccounted for.

Chapter 1. The Promise and the Problems

In this chapter, I provide some background, introduce some terminology and set the stage for the argument of subsequent chapters. In the **Introduction** to this chapter, I show how two competing conceptions of justice, Justice as Reciprocity and Justice as Benevolence, both appeal to everyday notions of right and wrong, but appear to provide conflicting recommendations as to what we should do. In **Section 1** I briefly explain how 'deontological liberalism' is to be understood for the purpose of this thesis and in **Section 2** I put forward an account of classical act-utilitarianism and explain how it is unable to accommodate 'three tenets of deontological liberalism' that possess much intuitive appeal and that deontological liberalism *is* able to accommodate. I then turn in **Section 3** to a discussion of the merits and demerits of competing conceptions of justice to provide some background for the two that are the primary focus of this thesis; Justice as Reciprocity and Justice as Benevolence. I also explain Rawls's conception of Justice as Reciprocity in more detail; my interpretation is based on and coincides with one given by Allan Gibbard.

Section 4 sets out the defining features of Justice as Reciprocity. The argument of this thesis will revolve around my attempts to refute Rawls's efforts to show that utilitarian conceptions of justice are incompatible with Justice as Reciprocity as defined by those three features.

In **Section 5** I describe utilitarian liberalism with particular reference to the arguments of its greatest historical exponent, John Stuart Mill. The purpose of this is to present an alternative philosophical foundation for liberalism in contrast to the 'deontological liberalism' of Section 1 which is espoused by Rawls and to point out its drawbacks. In the brief **Section 6** I put my view that the attractiveness of Rawls's approach to distributive justice lies largely in its promise to provide a foundation for deontological liberalism with Justice as Reciprocity.

Section 7 describes Rawls's particular theory of Justice as Reciprocity, Justice as Fairness, in some detail and **Section 8** describes Rawls's conception of Justice as Benevolence.

Section 9 outlines the problems that I maintain that Rawls was never able to surmount and **Section 10** explains their significance for the rest of the thesis.

Introduction

1.0.1 I first approached John Rawls' *A Theory of Justice* with two worries in mind. I was sympathetic to utilitarianism and had been unconvinced by some of the criticisms I had read of it in various other writings that lay more in the field of moral than political philosophy.[1] And utilitarianism seemed to capture my intuitions about the requirements of benevolence. So I believed that insofar as I had the right to be benevolent, and insofar as I had a duty to be benevolent, it would be right to maximize happiness even if that meant providing more benefits to better off people at the cost of lesser benefits to worse off people. But I had the worry that justice in distribution required people to fairly repay those they owed for the advantages that they received from them, and that this requirement would conflict with utilitarianism, however adequate utilitarianism might be as a theory of how best to be benevolent.

1.0.2 The second worry was that utilitarianism seemed to conflict with the idea that we should have the right to live our own lives the way we chose, so long as we respected other people's freedom. The American poet Robert Frost once said 'I hold it to be the inalienable right of anybody to go to hell in his own way,' and I sympathized with that view, even while holding that it wouldn't be particularly laudable to use your inalienable right in that way.[2] The problem would be how to set limits on the exercise of that right. Going to hell in one's own way could mean driving as fast as you could in the wrong direction down the M4. But that would very likely take a number of people with you, and surely you didn't

[1] One such writing was Bernard Williams' influential *A Critique of Utilitarianism* (1973)

[2] Robert Frost would arguably have more appropriately referred to an 'inviolable' rather than an 'inalienable' right and my sympathies should have been with that revision of his view. 'Inalienable rights' are rights the possessor of which is not at liberty to divest herself of. 'Inviolable rights' are rights which place limits on the legitimate actions of others. Frost, presumably, intended to say that people having the right to go to hell in their own way meant that others should not interfere with a person's right to decide to go to hell.

have the right to do that.

1.0.3 To illustrate the first of these worries with a simple example. I have an unexpected lottery win of £100. A friend of mine recently fixed my van for free as a favour when I was broke, when normally I'd have paid her £100. Now I know that she's quite hard up (but not destitute) and I set out to give her the money. On the way I am waylaid by a charity fundraiser for UNICEF. He persuades me that by donating £100 to their cause, UNICEF can buy enough vaccines to save 100 children's lives in Africa. The expected benefits of my giving the money to the charity appear, then, to exceed the expected benefits of returning the favour to my friend. So my benevolent instincts favour giving the money to the fundraiser from UNICEF. But I am worried that justice requires me to give the money to my friend. I *owe* her the money in return for the work she carried out on my van. However much good I might do for children by giving to the charity, they have done nothing to benefit me and consequently I owe them nothing. Furthermore, *justice* requires me to repay debts to the people I owe before I start thinking about the amount of good I could bring about.

1.0.4 The second worry can be explained as the worry that utilitarianism would limit my freedom to go to hell in my own way (traffic regulations can surely be justified by utilitarian cost-benefit analysis), at too high a cost. It doesn't appear to give me any right to do anything at all apart from promote the greatest happiness of the greatest number. To take a statement of the utilitarian creed from one of its most widely known and uncompromising contemporary advocates, Peter Singer

> In its simplest, classical form, utilitarianism is the theory that an act is right only if it does at least as much to increase happiness and reduce misery, for all those affected by it, as any possible alternative act.[1]

So according to this code, it wouldn't just be my right to give the £100 to UNICEF, it would be my duty. Utilitarianism doesn't appear to allow people *any* freedom to live their

[1] Singer 1981 p.64

lives the way they choose.[1]

1.0.5 Underlying the first worry is the suggestion that there may be another source of moral motivation than benevolence that grounds our fundamental obligations of justice in distribution: *reciprocity*, which for now can be defined just as the requirement that we give fair repayment for our advantages to those who provide them. I think the example above serves to illustrate the difference between the two sources of moral motivation well enough, though it should not be taken to provide an illustration of what the actual demands of reciprocity might turn out to be, as a moment's reflection should show. Suppose we accept that the motive of reciprocity as just described grounds our fundamental obligations of justice in distribution. Following various commentators, most notably Allan Gibbard and John Rawls himself, I shall call this conception of distributive justice, **Justice as Reciprocity**. The example is set against background institutions of society which are all open to question. It may turn out, when the demands of Justice as Reciprocity are fully cashed out, that the mechanic had no right to own the garage I took my van to, or that I had no right to the van or the money in my pocket, or all, some, or none of the above. But given the very different nature of the two sources of moral motivation under consideration, whatever the requirements of Justice as Reciprocity turn out to be, when fully cashed out, it seems likely that they will conflict with utilitarianism when utilitarianism is assumed to depend on the motive of benevolence for its foundation.

1 Deontological liberalism

1.1.1 The second worry would be solved by the existence of 'inviolable rights' that protect the individual's rights to live the life they choose from the demands of utilitarianism while also imposing bounds on acceptable behaviour to live the life of their choice so long as they don't violate others' right to do the same. So I should have the freedom not to give every last penny to good causes while not having the freedom to drive the wrong way on the motorway. And so should everybody else. This position is **deontological liberalism**

[1] Samuel Scheffler argued that consequentialism could, and should, be adapted so as to allow moral agents more freedom through an 'agent-centred prerogative' in *The Rejection of Consequentialism* (1994). More recently the criticism that utilitarianism denies people sufficient freedom has been upheld by Peter Vallentyne in his essay, 'Against Maximizing Act-Consequentialism (2006).

and I hope its appeal is already obvious from the account I've given of it above. Deontological liberalism I take to be defined by the acceptance of one or more of three tenets. The first, which I shall call the **core tenet of deontological liberalism,** is that people have the right *not* to maximize the aggregate sum of advantages, however advantages may be defined. The second is that the rights that people have impose some limits on what people can rightly do to others. Robert Nozick offers a useful account of this second tenet of deontological liberalism in his book *Anarchy, State and Utopia,* '[a] line (or hyper-plane) circumscribes an area in moral space around an individual. Locke holds that this line is determined by an individual's natural rights, which limit the action of others. Non-Lockeans view other considerations as setting the position and contour of this line.'[1]

1.1.2 It is useful here to distinguish more clearly between the core tenet of deontological liberalism and the second tenet of deontological liberalism. The core tenet of deontological liberalism bestows on people rights that were famously categorized by the legal theorist Wesley Hohfeld as 'liberty-rights'. A 'liberty-right' accords a person the right to do something without placing others under any correlative duty to restrain their behaviour in respect for that right (without limiting the actions of others, in Nozick's words). In the present context, this would mean that that those who have the liberty not to maximize utility are under no obligation to maximize utility. But it would not place anyone else under any obligation to respect their freedom to not maximize utility. If there were people strong enough to force others to maximize utility against their will, the strong would have the right to do so. What would be needed for the weak to be protected against the strong in this case are Hohfeld's 'claim-rights' that would be bestowed by the second tenet of deontological liberalism. A 'claim-right' not only gives people a right to something, it places others under an obligation to respect that right (in the words of Nozick it *would*

[1] Nozick 1974 p.57. Robert Nozick counts as a Lockean whereas Rawls, I think, would be a non-Lockean in Nozick's terms as the deontological rights he defends are not 'natural' ones, but ones that are derived from a hypothetical social contract. It is worth noting that on more than one occasion Rawls appears to subscribe to a different definition of 'natural right'. So, in *A Theory of Justice* (henceforth *Theory*) he writes that '[e]ach member of society is thought to have an inviolability on justice, or as some say, on natural right, which even the welfare of everyone else cannot override' and he maintains that his hypothetical contract can account for this 'common sense conviction' of justice (Rawls *T of J Rev* 1999 p. 24-25).

'limit the action of others.')[1]

1.1.3 This distinction is important as my argument of Chapter 4 that 'reciprocal classical utilitarianism' would only respect the core tenet of deontological liberalism involves demonstrating that it would only grant 'liberty-rights' and not 'claim-rights.'

1.1.4 The third tenet of deontological liberalism is that the rules of justice that define peoples' rights should be, at least in some way, neutral between different conceptions of the good.

2 Utilitarianism: Act and Reciprocal

1.2.1 Utilitarianism, as it is commonly understood, is classical act utilitarianism. Classical act utilitarianism is unable to accommodate any of the three tenets of deontological liberalism. The statement from Peter Singer above (§1.0.4) provides, I think, a good statement of that form of utilitarianism and can be used to explain just why it is unable to accommodate of these three tenets. The core tenet of deontological liberalism holds that people should have the right not to maximize the sum of advantages. But all are obliged, according to Singer's statement, to perform only those acts that do at least as much to increase happiness and reduce misery as any possible alternative act. Then all are obliged to perform those acts that maximize the 'sum of advantages', on classical utilitarianism's understanding of the notion of 'maximizing the sum of advantages'. It follows that classical utilitarianism does not grant me the right to not perform the acts that maximize the sum of advantages. So classical act utilitarianism is unable to accommodate the core tenet of deontological liberalism which holds that I do have the right not to maximize the sum of advantages.

1.2.2 It is also unable to accommodate the second tenet. If I am *obliged* to perform that act which maximizes advantages, then I am certainly *permitted* to perform that act, and so

[1] My account of Hohfeld's distinction between 'liberty-rights' and 'claim-rights' follows that given by Eleanor Curran in 'Blinded by the Light of Hohfeld: Hobbes's Notion of Liberty' (Curran 2010 pp. 100-102)

no limits can be placed on my right to perform that act. But the second tenet of deontological liberalism, as the quotation of Nozick's illustrates, claims precisely that people have rights which would limit what I could permissibly do in the pursuit of maximizing the sum of advantages. A variant of a well-known example against utilitarianism imagines a doctor in a hospital who has four patients who will shortly die unless they receive an organ transplant. One needs a new liver, the second a new lung, the third a new pancreas and the last a new heart. As chance would have it, there's a little old lady fast asleep on a bench around the back of the hospital. Concerned that she need needs friends and relatives to come and collect her, the doctor goes through her handbag in search of an address book. He doesn't find one, but does find an organ donor card, certifying that all her organs had been checked the week before and are in phenomenal condition for someone of such advanced years. He also happens upon a copy of her will which declares that since she has no friends and relatives she intends to leave what little money she has to the International Charity for Animal Welfare (ICAW): a charity which spends the bulk of its revenue on manufacturing toy animals to give to its donors as rewards for their donations. A quick calculation persuades the doctor that the utility maximizing action would be to suffocate the old lady, then have her found and wheeled in for her organs to be harvested. But the second tenet of deontological liberalism claims that she has some kind of right, perhaps the right to life, that would be violated by the contemplated homicide and that thereby renders such an act morally impermissible. Because act utilitarianism holds that utility maximizing acts are permissible it cannot accommodate the second deontological precept of justice, which sets limits on what we may permissibly do to others in the name of the greater good.

1.2.3 Utilitarianism is also unable to accommodate the third tenet of deontological liberalism. The term 'conception of the good' referred to above (§1.1.3) can be thought of as a choice of a particular kind of life. It is convenient here to exemplify the idea of 'conception of the good' rather than to define it precisely. The dominant forms of religion all prescribe different conceptions of the good to their followers by directing them to act in different ways. Muslims and Jews are directed to not eat pork and Hindus are directed to not eat beef. The person who wants to go to hell in his own way has a particular conception of the good, as does a classical act utilitarian who accepts that we are obliged to perform

only those acts which do at least as much to increase happiness and reduce misery as any other available act. And there lies the problem. If we are obliged to act in accordance with utilitarian criteria, then there is no room for any competing criteria to determine how we may behave. We should eat pork or beef if and only if doing so maximizes the sum of advantages. Thus, the doctrine of classical act utilitarianism allows no room to respect the right of Hindus to not eat beef and Muslims to not eat pork if they choose to do so on religious grounds. Classical act utilitarianism specifies that there is only one conception of the good that may legitimately order our lives: classical act utilitarianism. It is not neutral between itself and other conceptions of the good.

1.2.4 The variant of utilitarianism considered so far has been classical act utilitarianism, which defines the advantages to be maximized in terms of pleasure minus pain. Those utilitarian theories which define the good to be promoted in this way are known as 'hedonistic.' Classical act utilitarianism not only defines the advantages to be maximized in terms of happiness, but also aims to maximize the total sum of such advantages. Another possibility would be to aim at maximizing the average level of utility in society. This alternative is 'average utilitarianism' and is the variant Rawls considers to be the main rival to his principles of justice. There are also other ways in which 'advantages' may be defined by different variants of utilitarianism. Alternatives to hedonistic utilitarianism include those variants which define the advantages to be maximized in terms of preference satisfaction or something similar. Despite their considerable differences, all the versions of utilitarianism considered by my account so far could embrace the notion that all moral agents are morally obligated to perform those acts that maximize, or could be expected to maximize, the sum of advantages, however this may be defined. They would then, by definition, all be variants of **act utilitarianism.** But there are other variants of utilitarianism that haven't yet been considered, which are often described as 'rule' or 'indirect' utilitarianism. These do not require the moral agent to always perform those acts that maximize, or could be expected to maximize, the good. Instead, they direct people to act in accordance with rules that derive their ultimate justification from some kind of idea

that they are the rules with the greatest propensity for maximizing the good. [1]

1.2.5 For the purpose of this thesis I shall set rule (or indirect) utilitarianism aside and be concerned with act utilitarianism. I shall also be arguing on behalf of the classical principle of utility, so my arguments should often be read as assuming 'utilitarianism' to refer to 'classical act utilitarianism'. But I believe that much of my argument on behalf of classical act utilitarianism might be coherently extended in support of other variants of utilitarianism, including variants of rule utilitarianism, though it is beyond the remit of this thesis to extend them so far. It is, however, very important to my argument to introduce the new variant of utilitarianism of my own devising, that I have referred to as '**reciprocal classical utilitarianism'. Reciprocal classical utilitarianism** acknowledges the classical principle of utility as the predominant principle of distribution for society, but allows that some individuals may be exempt from the obligation to maximize the good. This distinction is necessary, I shall argue, as the logic of **Justice as Reciprocity** points to those for whom cooperation with others holds sufficiently dismal prospects to be relieved from a duty to cooperate. I shall argue that while the charge that Rawls makes, that classical utilitarianism is unjust, because it involves a violation of rights, be made to stick, the same charge would not be upheld against reciprocal utilitarianism.

1.2.6 In contrast to classical act utilitarianism, reciprocal classical utilitarianism can accommodate the core tenet of deontological liberalism.[2] However, my argument will not provide much consolation for deontological liberals as they are typically committed to the other two tenets of deontological liberalism as well. And reciprocal utilitarianism is, I shall show, unable to accommodate the other two. This demonstration won't be fully completed until Chapter 4, but I can here give an indication of its implications. In contrast to act utilitarianism, reciprocal utilitarianism would relieve people of a duty to perform utility maximizing actions if performing those actions would jeopardize their prospect of having a life worth living. But, by not upholding the second tenet of deontological liberalism,

[1] A comprehensive overview of the different variants of utilitarianism, including more obscure ones such as 'scalar utilitarianism', is given by Julia Driver in *Consequentialism* (2012).
[2] I also suspect that other variants of utilitarianism, including rule utilitarian variants, would face insurmountable difficulties in accommodating the core tenet of deontological liberalism though it is beyond the remit of this thesis to prove this.

reciprocal classical utilitarianism would always permit someone to perform utility maximizing actions. So in the context of the anti-utilitarian example given above (§1.2.2), reciprocal classical utilitarianism would permit the doctor to murder the old lady and harvest her organs, without construing such an action as a violation of her rights. It would also not be neutral between conceptions of the good. Insofar as someone has an obligation to act reciprocally, that is, if they are not exempted from the obligation to perform the action in question on the grounds that doing so would jeopardize their prospect of having a life worth living, they are obliged to maximize utility at the expense of promoting any other value.

3 Competing conceptions of justice

1.3.1 I think it is fair to speculate that Rawls initially approached the subject of distributive justice with similarly conflicting worries to those of mine that I described above (§1.0.1 - §1.0.5).[1] The way he resolved them, on the reading of his theory that I shall uphold for the purpose of this thesis, was to put forward a theory of Justice as Reciprocity. According to Rawls's theory of Justice as Reciprocity, the rules that people should follow in order to provide fair repayment for their advantages to those who provided them would be rules that also instantiated the three tenets of deontological liberalism. The particular theory of Justice as Reciprocity Rawls put forward, he named Justice as Fairness, and it relied on the device of a hypothetical contract to construct principles suitable for the conception of Justice as Reciprocity. In the next section I shall describe Rawls's theory of Justice as Fairness in more detail. But here I clarify the broader conception of Justice as Reciprocity by setting it out alongside rival conceptions of justice. Doing so will enable me to flesh out the conception of Justice as Reciprocity in more detail, and compare some of its advantages and disadvantages with those of its rivals, including, most pertinently for this thesis, the conception of Justice as Benevolence.

[1] This may sound surprising to those who are only familiar with his works from 'Justice as Fairness' onwards, as those works display little sympathy with utilitarianism as a theory of distributive justice, though they do show some admiration for its method and its exponents. But Rawls's early essay 'Two Concepts of Rules' (1955) is best read as an attempt to put forward a kind of rule-utilitarianism and is certainly sympathetic to the general aims of utilitarianism.

1.3.2 The easiest way for me to explain some of the important features of different conceptions of justice is against the backdrop of a debate between Allan Gibbard and Brian Barry over how Rawls's overarching philosophical project should best be understood. My sympathies in this debate lie entirely with Gibbard's suggestion that Rawls's theory is essentially a version of Justice as Reciprocity, and I shall start by giving an account of Gibbard's position in that debate.

1.3.3 Gibbard argued that Rawls should be understood as putting forward a theory of Justice as Reciprocity in his article 'Constructing Justice', a review of Brian Barry's book *Theories of Justice.* To understand Gibbard's position, it will help to first have an idea of the main theme of Barry's *Theories of Justice. Theories of Justice* examines Rawls's particular theory of justice in the light of what Barry takes to be the two main strands of thought on distributive justice, throughout the history of philosophy. The first strand is what Gibbard, following Barry, calls 'Justice as Mutual Advantage'. This takes the motive to act justly to be an egoistic one: people should act justly because it is in their own interest to do so. The historical name most associated with this approach towards justice is the seventeenth century English philosopher Thomas Hobbes, who shall receive more attention in this chapter and in Chapter 4, but elements of the conception of Justice as Mutual Advantage are found in Plato's *Republic,* relates Barry.[1] Much more recently, a contemporary of Rawls, David Gauthier, put forward a Hobbesian theory of justice in his book *Morals by Agreement* in 1986.

1.3.4 Justice as Impartiality has a vaguer history, according to Barry, but its 'general approach, which calls on people to detach themselves from their own contingently given positions and take up an impartial standpoint…is a product of the enlightenment, and everyone who follows it acknowledges a debt to Kant'.[2] The defining feature of Justice as Impartiality, Barry stipulates, is its adoption of the motive for behaving justly as 'the desire to act in accordance with principles that could not reasonably be rejected by people seeking an agreement under conditions free from morally irrelevant bargaining advantages and

[1] Barry 1989 pp. 3 - 9
[2] Barry 1989 p. 8

disadvantages.'[1] Barry interprets Rawls as being torn between these two approaches to justice, uneasily combining elements of both in *A Theory of Justice*.

1.3.5 However, Gibbard suggests, as an alternative interpretation of Rawls's project to Barry's, that 'Rawls long ago seemed to have his eye on a third perch: one he called Justice as Reciprocity'.[2] This conception of justice should be distinguished from Justice as Mutual Advantage by its espousal of an alternative motive to be just. As Gibbard describes this alternative motive

Passage 1a: ***(CJ)***

> If I return favor for favor, I may be doing so in pursuit of my own advantage, as a means to keep the favors rolling. My motivation might, however, be more intrinsically reciprocal: I might be decent to him because he has been decent to me. I might prefer treating another well who has treated me well, even if he has no power to affect me. We tip for service in strange restaurants.[3]

1.3.6 The motive Gibbard describes as central to Justice as Reciprocity is essentially the same as the motive I described in my example above (§1.0.3 - §1.0.5). But where my example was designed to illustrate an intuitive conflict between the demands of benevolence and reciprocity, Gibbard's example nicely illustrates the distinction between acting egoistically, the motive behind the conception of Justice as Mutual Advantage, and acting reciprocally. We have a self-interested motive for tipping in a local restaurant because, if we don't, the waiter might spit in our food on our next visit. But we have no self-interested motive for tipping in a restaurant that we won't return to. To tip in these circumstances is to draw on another source of motivation: reciprocity. Justice as Reciprocity takes that source of motivation to be our motive to be just.

---

[1] Barry 1989 p. 8. Barry takes his description of the motive underlying Justice as Impartiality from Thomas Scanlon's 'Contractualism and Utilitarianism'. But Barry's interpretation of Scanlon as a theorist of Justice as Impartiality is contentious, so I have not referred to Scanlon as a theorist of Justice as Impartiality. I am grateful to Mike Otsuka for pointing the contentiousness of Barry's interpretation of Scanlon out to me.

[2] Gibbard 1991 p. 266

[3] Gibbard 1991 p. 266.

1.3.7 Despite the different 'motives to be just' underlying Justice as Reciprocity and Justice as Mutual Advantage, advocates of the two competing conceptions of justice agree on the point that society should be conceived of as *a cooperative venture for mutual advantage*. This agreement over the fundamental conception of society appears to have led to an agreement between the main contemporary proponents of the rival approaches, John Rawls and David Gauthier, on the scope both of persons who have any obligation of justice, and of those to whom it should be extended. My reciprocal obligation to the waiter in the strange restaurant, if I have one, says that because *they* have benefited *me*, *I* should benefit *them.* The beneficiary of the favour has an obligation to return a favour to his benefactor. But an obligation of reciprocity or mutual advantage would not extend to providing benefits for people in developing countries who have done nothing to benefit me, or for some of the congenitally disabled, who have not done, and cannot do, anything to benefit me.[1] And here lies a simple difference between Justice as Reciprocity and Justice

[1] I am here passing swiftly over two complicated issues. The first is that of whether Justice as Reciprocity excludes the congenitally disabled from its scope. My interpretation of Rawls here follows that of Barry who writes in *Justice as Impartiality* that 'the grim logic of justice as reciprocity excludes them [i.e. the congenitally disabled] from its scope'. (Barry 1995 p.60). Barry's evidence for Rawls's exclusion of the congenitally disabled is that '[i]n *Political Liberalism*, Rawls says that among the problems left over by his theory, which assumes that 'persons are normal and cooperating members of society over a complete life' is the question of what is owed to those who fail to meet this condition, either temporarily (from illness and accident) or permanently' (p.21). He adds that he thinks the theory 'yields reasonable answers…to part of [this question], to the problem of providing for what we may call normal health care.' Barry reads this as an 'an implicit admission that Rawls cannot, any more than could Gauthier…accommodate the idea that justice demands support of the congenitally disabled.' (Barry 1995 p.60). Rawls did not widen the scope of his theory to include the congenitally disabled in *Justice as Fairness: A Restatement (2001)* where he reaffirmed his view of 'political society…as a fair system of social cooperation…where those engaged in cooperation are viewed as free and equal citizens and *normal cooperating* members of society over a complete life.' (Rawls 2001 p. 4 my italics). So my interpretation of Rawls's theory is in accordance with Barry's in holding it to exclude the (sufficiently) congenitally disabled from either having obligations of reciprocity themselves or being those to whom obligations of reciprocity are owed. I shall, however, suggest in Chapter 4 that the logic of Justice as Reciprocity may extend to including some of the congenitally disabled within the scope of those to whom obligations of reciprocity are owed. The second contentious issue is whether the logic Justice as Mutual Advantage is extended too far in including within the scope of persons to whom obligations of mutual advantage are owed those who have no power to harm us if we don't provide them with benefits. Gauthier's theory of mutual advantage concurs with Rawls's theory in extending the scope of persons to whom obligations of justice are owed to include all normal cooperating members of society. But many of those, such as the waiter in the strange restaurant, would have no power to harm us if we did not fulfil our obligations to them, so a question arises as to why we should be motivated by self-interest to fulfil such obligations. Gauthier's answer is, roughly, that rational self-interested behaviour should be evaluated at the level

as Impartiality, at least as Barry conceives of it. Justice as Impartiality, as described by Barry in *Theories of Justice* and its sequel *Justice as Impartiality*, extends the scope of my obligations of justice to include anyone who might be affected by an agent's behaviour. People in developing nations who could benefit from my making a small donation to UNICEF might be affected by my behaviour as whether or not I make that donation could make a difference to their lives. So they should, arguably, be included in the class of people who might reasonably reject any principle that exempted me from making such donations. Justice as Impartiality, as put forward by Barry, would include more people within its scope than Justice as Reciprocity.

1.3.8 However, as Gibbard remarked in 'Constructing Justice', 'Justice as Impartiality may be too vague to be a clear, distinct alternative' to Justice as Mutual Advantage or Justice as Reciprocity.[1] He exemplified this point by showing how justice's requirement of equal treatment, an element of Rawls's theory that Barry admires and interprets as an element of Justice as Impartiality within Rawls's theory, could be reconciled with utilitarianism. As Gibbard puts it '[u]tilitarianism treats everyone equally, in the sense that every person is told that his interests will be overridden only when otherwise others would have to forgo a greater interest.'[2] Gibbard then questions the need for Justice as Impartiality to go beyond the equal treatment that is compatible with utilitarianism, as Barry asserts that it should, in favour of a more egalitarian theory of justice. So Gibbard effectively asserted that the motive to act impartially was neutral between utilitarianism and the more egalitarian approach to distributive justice favoured by Barry and Rawls.

1.3.9 Barry responded by subsequently putting a case for Justice as Impartiality's

---

of disposition rather than action. In terms of this example, if I were the kind of person who wouldn't leave a tip the waiter would be aware of this from the start and not provide me with good service in the first place, so it is not in my self-interest to be the kind of person who would not leave a tip. Gibbard offers some convincing (to my mind) objections to this aspect of Gauthier's theory in 'Constructing Justice' as does Parfit in *Reasons and Persons*. (1984 pp. 16 - 23) For the purpose of my thesis I can afford not to enter any further into this debate, as I touch on the conception of Justice as Mutual Advantage only to provide some insight into the conceptions of Justice as Reciprocity and Justice as Benevolence. A proper evaluation of the viability of Justice as Mutual Advantage is beyond the remit of this thesis.

[1] Gibbard 1991 p. 266

[2] Gibbard 1991 pp. 277-278

incompatibility with utilitarianism in his book *Justice as Impartiality,* based on the idea that the principle of utility is a principle that could be reasonably rejected by those who would be worst off under its application, and so would not meet Justice as Impartiality's motivational requirement described above (§1.3.3).[1] I find Barry's case unconvincing, but do not have space to argue against it here.[2] But I will offer one consideration in favour of the rejection of Justice as Impartiality as an appealing alternative conception of justice, at least as Justice as Impartiality is put forward by Barry. One of the advantages of Justice as Impartiality over Justice as Mutual Advantage appeared to be its ability to extend the scope of those to whom obligations of justice are owed to 'third parties' such as the congenitally disabled and future generations. As I remarked above (§1.3.5), Justice as Impartiality, Barry maintains, would include these third parties within its scope because they are people who could 'reasonably reject' principles that neglected them. Many people intuitively feel that we have a moral duty to provide for those who are unable to provide for themselves and Barry's theory of Justice as Impartiality purports to offer one way to underwrite this intuition.

1.3.10 But would Barry's Justice as Impartiality really cover the 'third parties' who are future generations? If we take the motivational requirement of Justice as Impartiality literally, they are arguably beyond its scope as well. Future generations could not reasonably reject our refusal to bring them into the world, because in order to reject it they would have to exist, or at least be identifiable as people who are (at least probably) going to come into existence. Barry's Justice as Impartiality looks as though it may be unable to underwrite a strong intuition that many people share: we have a duty to preserve the planet for the sake of future generations.

1.3.11 In that regard Justice as Benevolence has a definite advantage. It could, and should, extend its scope to include future generations. If our motive to be just directs us to produce as much good as possible then we may have obligations to bring happy people into the world, regardless of whether we owe them in return for the benefits they provide for us

[1] See Barry *Justice as Impartiality* 1995 pp 61 - 67
[2] I put forward some considerations against in in Chapter 2 (§§ 2.5.2)

(which they don't) as Justice as Mutual Advantage would require, and regardless of whether they could reasonably reject a principle that didn't bring them into the world (they probably couldn't) as Barry's Justice as Impartiality would require. Justice as Benevolence would have the widest scope of all the conceptions of justice described in this section, extending to the congenitally disabled, future generations, non-human animals and even extra-terrestrials (if they do, or could, exist).

1.3.12 Unfortunately, despite what I take to be its very considerable advantage in holding out a prospect of underwriting an obligation to future generations, Justice as Benevolence doesn't strike me as a very plausible conception of justice. I shall say a bit more about this in Section 6 of this chapter.

1.3.13 It would be too much of a distraction from the main aims of this thesis to evaluate the conceptions of Justice as Impartiality and Justice as Mutual Advantage in more detail. The point of raising them was primarily to provide some context for the conceptions of Justice as Benevolence and Justice as Reciprocity. But I shall end this section with a couple of remarks as to why I think they would ultimately prove unviable.

1.3.14 Justice as Mutual Advantage would prove unviable because its motive to be just does not provide an adequate motive for just behaviour, as it is normally understood, as requiring a constraint on the pursuit of self-interest. Gauthier's attempts to sustain the theory by applying the notion of rational self-interest to disposition rather than action have, in my opinion, been effectively rebutted by Derek Parfit in *Reasons and Persons.*[1]

1.3.15 Barry's Justice as Impartiality seems to me, as it did to Gibbard, to be an unstable compromise. I find it hard to see a rationale for its contractualist motivational requirement once the conception of society as a cooperative venture for mutual advantage is dispensed with, as it is, by Barry. And it has the disadvantage already considered above of being at least unclear in its support for obligations to future generations. Justice as Benevolence, as remarked above, shares Justice as Impartiality's commitment to equal treatment without

[1] See Parfit 1984 pp. 17 - 23

being saddled with its contractualist motivational requirement.

1.3.16 So the contest that I am concerned with in this thesis is between Justice as Benevolence and Justice as Reciprocity, with Justice as Mutual Advantage and Justice as Impartiality set to one side. A couple more remarks are required about how I shall understand Rawls's conception of Justice as Reciprocity in this thesis and once again I follow Gibbard's account from 'Constructing Justice.'

1.3.17 As Gibbard relates

Passage 1b *(CJ)*

> Rawls proposes that justice is fairness in exchange, but on a grand scale: it is fairness in the terms governing a society-wide system of reciprocity. The system consists in each person's supporting a basic social structure and drawing benefits from it. The citizen of a well-ordered society is motivated to return benefits fairly, and this general motivation becomes a motivation to conform to the rules of a social structure he considers fair.[1]

Gibbard fleshes out this account of Rawls's theory of Justice as Reciprocity by going on to suggest that 'the prime question Rawls addresses' is

Passage 1c *(CJ)*

> "Why limit myself in pursuit of my own advantage?" This is a question that can be asked also by a well-off person: he has much but why not go for more? Rawls, in effect, gives this answer: "You have what you have only because others constrain themselves, in ways that make for a fair cooperative venture for mutual advantage. Constrain yourself by these rules in return, and you give them fair return for what they give you." Whether this answer moves a person depends on his sentiments of fair reciprocity.[2]

4. The defining features of Justice as Reciprocity

[1] Gibbard 1991 p. 265
[2] Gibbard 1991 p. 269

1.4.1 **Passage 1c *(CJ)*** I take to define the essence of the conception of Justice as Reciprocity and it sets the requirement that principles of justice have to meet, when society is conceived as a cooperative venture for mutual advantage, in order to be suited to the conception of Justice as Reciprocity. Three important points about it should be noted. The first is that Justice as Reciprocity requires *constraint* on the part of those who are required to act justly. I shall call this **the constraint requirement**. The second is that the reply Gibbard imagines Rawls might give makes reference to a *fair* cooperative venture for mutual advantage. The implication is that a necessary condition that must be met for any individual participating in society to have a duty of reciprocity, is that the rules governing society are fair. I shall refer to this as the **fairness condition,** and in the course of this thesis will argue that Rawls never produced a satisfactory stipulation of the fairness condition**.** The third is that the cooperative venture is for *mutual advantage.* This implies that a second necessary condition that must be met for any individual cooperating with the rules of society is that society affords them some prospect of 'advantage'. I shall call this the **mutual advantage condition.**[1] How 'advantage' should appropriately be defined for the conception of Justice as Reciprocity, is a difficult question, and another one that I shall maintain Rawls never satisfactorily resolved. It should be understood as advantage with respect to *the relevant situation of equal liberty*, whatever that might turn out to involve. The reason Rawls never managed to resolve the question of the **mutual advantage condition**, I shall argue, is because he never managed to come up with a relevant situation of equal liberty that yielded the results he wanted.[2]

1.4.2 My Preface concluded by claiming that a reasonable argument can be reconstructed that favours the classical principle of utility as a principle suited to the conception of Justice as Reciprocity. The features of the conception of Justice as Reciprocity that I have just highlighted are at the heart of my case, and also at the heart of Rawls's case that

[1] The justification for this term, as Chapter 3 will demonstrate, is to pick one that Rawls didn't use himself at various points to describe other conditions.

[2] In the first model of his theory it appeared to be the greatest liberty compatible with an equal liberty for all guaranteed by the first principle of justice (§§2.1.1 – 2.1.27). In the second model it appeared to be with respect to a vaguely described state of nature which wasn't as dismal as a 'Hobbesian' one (§§1.9.4 – 1.9.31). In model 1.5 it was an imaginary society where people were secure in their possession of constitutional liberties. In the third model it was a 'Hobbesian' state of nature of general egoism.

utilitarianism is not suited to the conception of Justice as Reciprocity. So I can set out here the broad outline of my argument, in relation to these features.

1.4.3 Rawls's case that the utilitarian conceptions of justice are unsuitable for the conception of Justice as Reciprocity revolved around his attempts to show that they could not meet either the fairness condition or the mutual advantage condition. As I describe in Chapter 3, his first model of Justice as Fairness looked able to support those contentions.[1] But he was forced to change that model when he realized that it didn't work, and never found a satisfactory substitute for the assumptions which, had they worked, would have sustained the charge that utilitarianism[2] would not meet the fairness or mutual advantage conditions. The result was that, according to the new assumptions of *Theory*, utilitarianism *would* meet the mutual advantage and fairness conditions. As I show in detail in Chapter 4 he devoted much energy and many words to trying to find some alternative mutual advantage condition that the two principles of justice would meet, but that utilitarianism would fail to meet. But, I shall maintain, he never succeeded in that endeavour. The only charge that he could, and did, continue to level against the principle of utility is that, in contrast to the principles of justice, it would be prepared to sacrifice the prospects of the less advantaged for the sake of greater advantages to the more advantaged.[3]

1.4.4 So the argument in this thesis revolves primarily around the question of whether reciprocal classical utilitarianism instantiates the three defining features of Justice as Reciprocity set out above (§1.4.1).

5 Utilitarian liberalism

1.5.1 Rawls's project should certainly not be understood as simply trying to prove that the conception of Justice as Reciprocity would repudiate utilitarianism. An equally important

[1] In fact, as I show in Chapter 2 (§§2.3.1 – 2.3.11) on the assumptions of the first model any principles of justice that met the mutual advantage condition would also meet the fairness condition and vice versa.
[2] To make for easier reading I shall sometimes use the term 'utilitarianism' to stand for all the various utilitarian conceptions of justice, bar reciprocal classical utilitarianism.
[3] This is the essence of the 'principle of reciprocity' referred to in several places in *Justice as Fairness: A Restatement* (e.g. Rawls 2001 p.77)

concern of his was to show that Justice as Reciprocity would provide a philosophical grounding for deontological liberalism, and he hoped to do this with his particular theory of Justice as Reciprocity: Justice as Fairness. I shall describe Justice as Fairness in some detail after first describing an alternative, and historically very influential, foundation for liberalism: utilitarian liberalism. Utilitarian liberalism would support the kind of liberal rights that Rawls was concerned to justify philosophically. But it would do so only tenuously and would rely on the problematic conception of Justice as Benevolence. Rawls evidently did not find this approach to the liberal values he held dear satisfactory and explicitly said as much in *Theory.*[1] The enormous influence Rawls's theories of justice and political liberalism can, I believe, be attributed to his championing of deontological liberalism, which appeals to philosophers who share Rawls's liberal values, and his disdain for the utilitarian justification of those values.[2] Many of those philosophers care little to nothing for Rawls's attempt to ground them in his theory of Justice as Fairness. However, in Chapter 4 I shall suggest that in the light of the failure of Rawls's project to provide a foundation for deontological liberalism through his theory of Justice as Reciprocity, utilitarianism may provide the strongest philosophical support for liberalism that is available. So I offer a sketch of utilitarian liberalism and the putative drawbacks of the utilitarian approach to liberalism here.

1.5.2 There is some irony in the fact that although Rawls's deontological liberalism and utilitarian liberalism have very different philosophical foundations, they would, in my opinion, be likely to prescribe quite similar policies in terms of both individual rights and economic distribution. The principle Rawls came up with to govern economic distribution,

[1] See *Theory* Chapter 6 (Rawls *T of J Rev*1999 p.25)

[2] Brian Barry provides an example of a philosopher who admires Rawls's liberal values but cares nothing for Justice as Reciprocity or the hypothetical contract device of Justice as Fairness. In *Justice as Impartiality* Barry writes: 'For the purpose of this book, I shall simply excise from Rawls's theory any reference to justice as reciprocity. If we do this we get a coherent theory of justice as impartiality.' (Barry 1995 p.60) Barry goes on, correctly in my view, interpret Rawls's theory of Justice as Fairness, as deriving its rationale from Justice as Reciprocity. But he then dismisses Justice as Fairness as an inadequate device as inadequate from the point of view of Justice as Reciprocity, remarking '[i]ts justification must come from its being the only way of producing the desired outcome. But then the redundancy of the whole contraction is patent.' (Barry 1995 p.61). I have much sympathy with Barry's criticism of Justice as Fairness from the perspective of Justice as Reciprocity and briefly offer a similar line of criticism in Chapter 3.

the difference principle, was one that would probably be met with considerable sympathy by most utilitarians. For the difference principle holds that the economic position of the worst off group in society should be made as well off as possible. Utilitarianism would also tend towards promoting the position of the worst off in society, at least in comparison with what it would have been in Western liberal societies such as the United States and the United Kingdom. This tendency of utilitarianism can be explained as follows: if we make the reasonable assumption that people have similar utility functions (utility could be defined as happiness or in terms of some other quantity as far as the present point is concerned) along with the other assumption of diminishing marginal utility, which holds that the utility of additional units of any good (including money) to a person declines after a certain point is reached, then utilitarianism would recommend at least rough equality, as that is where utility would be maximized.[1]

1.5.3 But, despite utilitarianism's natural bias towards economic equality, it is easy to understand how Rawls's *Theory* could hold a lot of appeal for left leaning liberals when it was first published, and ever since, as it appears to offer the hope of retaining what seems best about Western liberalism; its protection of people's right to choose how to live their lives free from coercion, while repudiating the worst; the economic inequality and poverty that 'free market society' could give rise to. And it should be noted that utilitarianism's bias towards economic equality outlined above depends on questionable contingent empirical assumptions. Should these turn out to be false then utilitarianism would have no such bias.[2]

---

[1] It is unfortunate, in my opinion, that often the first encounter undergraduate students of political philosophy have of utilitarianism is through reading *A Theory of Justice*, since *Theory* in several places depicts utility being maximized by bestowing advantages on the more advantages at the expense of the less advantaged. That utilitarianism would be prepared to contemplate giving greater advantages to the more advantaged in circumstances where the difference principle wouldn't is undeniable, but the impression that might be gained from the diagrams in *Theory* and *Justice as Fairness: A Restatement* is that this would be almost inevitable. See Rawls *T of J Rev* 1999 p.81 and Rawls 2001 p. 62.

[2] Robert Nozick (1974 p.41) raised the prospect of 'utility monsters' who gain much more than others from additional units of goods than others lose. If they existed, then utilitarianism would recommend redistribution from the less advantaged to these monsters, which seems unjust. But Allan Gibbard has offered an effective rebuttal of his point in *Reconciling Our Aims.* Gibbard observes that even if such monsters did, or could, exist, utilitarianism would pay regard to the 'incentive effects' and deprive such monsters as a means to discouraging others from emulating their example. See Gibbard 2008 p.72

1.5.4 If utilitarianism's bias towards equality depends on questionable empirical assumptions, then any bias it may have towards liberalism depends on assumptions that are far more questionable. The most famous political philosopher to uphold both liberalism and utilitarianism was John Stuart Mill (1808 - 1873). In *On Liberty* Mill argued strongly in favour of liberal rights on the grounds that granting people the right to live as they chose so long as they didn't harm others (and in many cases, such as in fair free market competition, allowing them the right to harm others) would promote a better and happier society in the long run, as the human race would naturally develop into a better, more utility maximizing species. As Mill eloquently described his position in *On Liberty*, the object of his essay was

Passage 1d (*On Liberty 1859)*

> to assert one very simple principle…that the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilized society, against his will, is to prevent harm to others…It is proper to state that I forego any advantage which could be derived to my argument from the idea of abstract right, as a thing independent of utility. I regard utility as the ultimate appeal on all ethical questions; but it must be utility in the largest sense, grounded on the permanent interests of a man as a progressive being.[1]

1.5.5 However, many critics from both left and right of the political spectrum have questioned whether it is true that 'the permissive society' would lead to a better and happier society in the long run as Mill hoped. And even if Mill's predictions were correct, there would be two more strong objections that could be levelled against his liberalism. The first is that Mill does not offer a sufficiently robust defence of rights: even those rights that he held were protected by the 'harm' or 'liberty' principle (as the principle of **Passage 1d (*On Liberty 1859)*** is variously described), he did not hold to be 'inviolable' but 'rights' which might justly be violated in exceptional circumstances. This compromise of Mill's often comes as something of a surprise to those who are primarily acquainted with Mill's *On Liberty*. But Mill quite explicitly asserted that such compromises might not only be

[1] Mill 2003 pp. 94-95.

morally permissible but required in *Utilitarianism* where he wrote that

Passage 1e (*Utilitarianism 1863)*

> …particular cases may occur in which some other social duty is so important, as to overrule any one of the general maxims of justice. Thus, to save a life, it may not only be allowable, but a duty, to steal, or take by force, the necessary food or medicine, or to kidnap, and compel to officiate the only qualified medical practitioner. In such cases, as we do not call anything justice which is not a virtue, we usually say, not that justice must give way to some other moral principle, but that what is just in ordinary cases is, by reason of that other principle, not just in the particular case. By this useful accommodation of language, the character of indefeasibility attributed to justice is kept up, and we are saved from the necessity of maintaining that there can be laudable injustice.[1]

1.5.6 The second objection to Mill's utilitarian liberalism is that insofar as it defended individual rights it did so for the wrong reasons. Mill defended the 'right of anybody to go to hell in his own way,' with great vigour (but not, as I have just pointed out, as inviolable) on the grounds that to do so would be good for society, as allowing people to go to hell in their own way would provide a salutary example to others.[2] This certainly doesn't seem to capture the spirit of the 'right' that Robert Frost believed in. To be sure, the 'right to go to hell in one's own way' might be a right that many would not want to defend, including some who could fairly be described as 'deontological liberals', for some who answer to that description would want to add a proviso that liberalism should only respect rational choices, and to choose to go to hell in one's own way is surely irrational. But deontological

[1] Mill 2003 p. 234. This quotation provides compelling evidence that Rawls's accusation that 'utilitarianism seeks to account for ['our convictions about the priority of justice'] as a socially useful illusion' (Rawls *T of J Rev* 1999 p. 25) could be fairly levelled at Mill. It raises other questions over the controversy over whether Mill should be regarded as primarily a 'rule' or 'act' utilitarian. Henry West (if I understand him correctly) seems to take Passage 1e as evidence that Mill might be interpreted as a rule-utilitarian on the grounds that it assumes that our behaviour should be governed by 'secondary rules' (the 'general maxims' of the passage) except when a direct appeal to the principle of utility is needed to decide between them (West 2004 pp. 87-88). But mightn't the fact that Mill allows arbitration by direct appeal to the principle of utility rather than insisting on ordering rules by priority be instead used as grounds for interpreting him as essentially an act-utilitarian? I prefer to interpret Mill as an act-utilitarian who (rightly) sets great store in the usefulness of abiding by strong rules of thumb, which isn't far from West's conclusion that 'Mill is neither a pure act-utilitarian nor a pure rule-utilitarian'. (Mill 2004 p. 95)

[2] Mill 2003 *On Liberty* Chapter 4.

liberals could go on to press their case against utilitarianism with more examples of rights whose defence would command much wider support, including the freedom to practice the religion of one's choice. Mill was very concerned to defend the freedom to practice religion, including Mormonism, despite his protestation that '[n]o one has a deeper disapprobation than [Mill] of this Mormon institution'.[1] But ultimately, this defence still rested on the long term interests of man as a progressive being. The deontological liberal would want to argue that whether Mormonism was or was not in the long term interests of man as a progressive being was beside the point. So long as the practitioners of Mormonism were genuinely consenting (Mill had in mind the practice of polygamy) it was none of anybody else's business.

1.5.7 The final objection to utilitarian liberalism is at the foundational level. As Rawls correctly observed, the utilitarian tradition has tended to rely on an assumption that we are naturally motivated, or have an obligation to, promote 'the good'. Its usual foundation is Justice as Benevolence. And this may seem to be an inadequate foundation, particularly if we accept Rawls's conception of society as a cooperative venture for mutual advantage.

6 The promise

1.6.1 I hope that the narrative of this chapter so far has conveyed something of what I take to be genuine advantages of the deontological liberalism that Rawls was concerned to uphold in comparison to the utilitarian liberalism of John Stuart Mill, in terms of their respective intuitive appeal. Deontological liberalism is able to accommodate the widely held belief that people have rights that can't (justly) be trampled over in pursuit of the greater good.

1.6.2 Rawls's theory also has, to my mind, great advantages over utilitarianism, as usually understood, in terms of its underlying motive to be just. Utilitarianism has seemed

[1] Mill 2003 p. 161. This might be an instance of Mill protesting too much. But, then again, it might not as he does go on to describe his aversion to it as 'a direct infraction of that [the liberty] principle, being a mere riveting of the chains of one half of the community [women], and an emancipation of the other half [men] from reciprocity of obligation towards them.'

to many to be an implausible ethical theory because of its inability to supply a compelling reason to be just. As explained above, Rawls took the underlying motive to be just to be one of reciprocity, of giving one's due to whom one owed one's due.

1.6.3 This may seem to many (including myself) to offer a better answer to the question of why we have a duty to act 'justly' than the idea that to do so would produce more good in the world. It is intuitively easier, I feel, to embrace the idea that one has a duty of *justice* to repay the people you owe in kind for what they have done for you than to embrace the idea that one has a duty of *justice* to promote the greatest good wherever that might be best achieved. That is not to rule out the possibility that there may not be other moral considerations that should direct us to promote the greatest good where we can, but if they do they are not duties of justice. Justice as Reciprocity is a more plausible conception of justice than Justice as Benevolence.[1]

1.6.4 I am also very sympathetic to the tenets of deontological liberalism. In the words of the song, we feel that 'it's my life and I'll do what I want'[2], captures the essence of what it is to have freedom or individual rights (though we would want to add a verse including caveats setting limits to our legitimate pursuit of what we want designed to protect the equal rights of others to do the same). 'It's my life and I'll do what I want – so long as doing so maximizes utility in the largest sense, grounded on the permanent interests of a man as a progressive being' doesn't sound such a compelling refrain.

1.6.5 The very real attraction of Rawls's project then, at least to my mind, was the promise it held out to underwrite strong and attractive intuitions about justice and the

[1] Disappointingly, I have no stronger argument for this position that I share with Rawls than that it feels intuitively sound. If no stronger argument is ultimately available then there is no objective basis for Justice as Reciprocity. It is of some consolation to note that, in my opinion, more ambitious recent attempts to found an objective morality on the concept of practical reason have also failed. David Gauthier's attempt to ground a contractualist moral theory by taking the rational pursuit of self-interest as foundational has, to my mind, been successfully defeated by objections such as those of Allan Gibbard and Derek Parfit mentioned earlier. And I am not convinced by Derek Parfit's attempts in *Reasons and Persons* to show that it the pursuit of self-interest is irrational, that could be read as an attempt to support the idea that Justice (or morality) as Benevolence is uniquely rational.

[2] Written by Roger Atkins and Carl D'Errico and first performed by The Animals (1965)

inviolability of rights with an equally strong and attractive account of their moral foundation.

1.6.6 Nevertheless, my argument is that attractive as Rawls's position may appear to be, it is ultimately unsustainable. I shall presently highlight some of the problems he faced by looking closely at the words he used to summarize the essence of his case that utilitarianism was incompatible with Justice as Reciprocity, early in *Theory*, and contrasting it with the very similar, but critically different, words he used to press the same charge against utilitarianism in his earlier essay, 'Distributive Justice' (1967).

1.6.7 But this comparison of passages will not make much sense without first introducing some more of the key elements of Rawls's own theory of **Justice as Fairness**, as it was presented in *A Theory of Justice,* and going into his understanding of utilitarianism in greater depth.

7 Justice as Fairness

1.7.1 As explained by the passages cited in my Preface, Rawls's starting point is that society should be conceived of as a cooperative venture for mutual advantage. This is a society of citizens who have personal interests that they are legitimately concerned to pursue. They could be described as non-altruistic in the sense that their primary concern may not be to promote the good of others, but to pursue these legitimate interests of their own.[1] But this does not mean that they are immoral. They respect other citizens' rights to pursue *their* legitimate interests and are prepared to restrain the pursuit of their own interests to give others a fair opportunity to pursue theirs. So this conception of society will need rules of justice to ensure that everyone gets a fair chance to pursue their own legitimate interests while respecting the rights of others to pursue theirs. But, in Rawls's theory, these rules of justice are not immediately obvious just from this conception of society. Instead, Rawls believed they needed to be constructed through the use of a carefully designed thought experiment. This is his conception of **justice as fairness.**

[1] This does not exclude the possibility that some citizens' primary 'personal' interest might be to promote the good of others.

Justice as Fairness holds that the correct principles to govern society conceived of as a cooperative venture for mutual advantage are whichever ones would be chosen in an appropriately specified hypothetical contractual situation. So to the end of discovering these, Rawls did exactly what Mill wrote would serve no purpose in my Epigraph from *On Liberty*; he invented a contract in order to deduce social obligations from it. Rawls's hypothetical contract model is rather complicated and a rough summary will suffice for the time being.

1.7.2 Rawls imagined representatives of the different social groups in society abstracted from their real world conditions. The circumstances they were to be imagined in he called 'the original position'. These representatives are shielded by a 'veil of ignorance' from knowledge of all the facts and circumstances that might unfairly bias their choice of principles, such as their natural abilities and talents, their class position and social status, the level of development of society and even which generation they belong to. They do, however, 'know the general facts about human society. They understand political affairs and the laws of human psychology. Indeed, the parties are presumed to know whatever general facts affect the choice of the principles of justice.'[1] They also know that they are in 'the circumstances of justice', which are, roughly, that the conditions they will find themselves in once the veil of ignorance is lifted won't be so harsh that cooperative schemes will break down, but also that resources aren't so abundant that there is no need for rules to decide how those resources should be allocated amongst competing claims. Importantly, the parties in the original position are to be conceived of as free and equal, rational and self-interested.

1.7.3 Rawls maintains that the parties in the original position would choose his two principles of justice. The first principle is the equal liberty principle, which stipulates that each person should have an equal right to the most extensive liberty compatible with a similar liberty for all.[2] 'Liberty' in *Theory* is defined in terms of a set of 'basic liberties';

[1] Rawls *T of J Rev* 1999 p. 119

[2] Rawls later changed the formulation of the equal liberty principle in response to criticisms by H.L.A Hart. His essay 'The Basic Liberties and their Priority' in *Political Liberalism* (1996) substitutes 'a fully adequate scheme' for *Theory's* 'the most extensive system.'

these include 'political liberty (the right to vote and to hold public office) and freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person, which includes freedom from psychological oppression and physical assault and dismemberment (integrity of the person); the right to hold personal property and freedom from arbitrary arrest and seizure as defined by the concept of the rule of law'.[1] Rawls's second principle of justice is divided into two parts. The first part holds that social and economic inequalities are to be arranged so that a) the position of the worst off group in society is maximized and b) the positions and offices to which these inequalities are attached are open to all. This part Rawls calls '**the difference principle'.** The second part is '**the equality of opportunity principle'**.

1.7.4 From the perspective of Justice as Reciprocity, the two principles of justice should be regarded as subordinate to Rawls's conception of Justice as Fairness. *Whichever* principles would be chosen by the parties in the original position would, according to Justice as Fairness, be those that should be rightly regarded as 'just', whether they are the two principles of justice, utilitarianism, or any of the other conceptions of justice available for consideration by the parties. It is only if, as Rawls maintains they would be, the two principles of justice emerge as the choice of the parties in the original position that they should ultimately earn the designation 'just' on the definition of 'just' given by Justice as Fairness. This point needs to be spelt out, as at points in *A Theory of Justice* Rawls inappropriately uses the term Justice as Fairness to designate the two principles of justice in contexts when it is the very question of which conception of justice the parties would choose which is at stake.[2] But early in *Theory* he is quite explicit that his principles of justice derive their "fairness" from the fact that they would be selected by the hypothetical contract. So he writes

[1] Rawls (*T of J Rev* 1999 p. 53) introduces this list with the words '[i]mportant among these are', implying that the list should not be viewed as exhaustive.
[2] I point out Rawls's inappropriate use of the term 'justice as fairness' in Chapter 3.

Passage 1f (*T of J Rev)*

> The original position is, one might say, the appropriate initial status quo, and thus the fundamental agreements reached in it are fair. This explains the propriety of the name "justice as fairness": it conveys the idea that the principles of justice are agreed to in an initial situation that is fair.[1]

1.7.5 The outline just given of Rawls's theory shows how it could be read as a theory of Justice as Reciprocity, with individuals behaving in accordance with the two principles of justice being motivated by reciprocity. Through compliance with the two principles of justice, co-operating members of society give fair return to others for the benefits that those others' compliance affords them.

1.7.6 So Rawls's two principles of justice may turn out to be the principles of justice most suited to the conception of Justice as Reciprocity. Whether they do or not depends on two arguments going through. The first is that the parties in the original position would pick the two principles of justice over any other conception of justice. The second is that Justice as Fairness is the right way to go about constructing principles suitable for the conception of Justice as Reciprocity. Both of these arguments can be questioned. The first, in particular, has come under considerable fire, even from critics who are sympathetic to Rawls's two principles.[2] The second has received less attention, but I shall subject it to some criticism of my own in the course of this thesis.

1.7.7 However, from what has been said so far, it may also be the case that the principle of utility, in either its average of classical version, may turn out to be the conception of justice most suited to the conception of Justice as Reciprocity.

8 The utilitarian conception of society

1.8.1 Rawls, in *Theory*, conceived of utilitarianism as starting from a very different point; he saw it as a theory of Justice as Benevolence. And as I remarked in the Preface; given the

[1] Rawls *T of J* 1999 p. 11
[2] For example Brian Barry in *Theories of Justice* (1989 pp. 214 – 215)

very different conceptions of Justice as Benevolence and Justice as Reciprocity, it seems, on the face of it, probable that they will yield different practical recommendations. In order to assess whether they would or not it is also necessary to explain how Rawls conceived that utilitarianism might be constructed. He sets out one possible foundation for utilitarianism thus

Passage 1g *(T of J Rev)*

> [t]he most natural way, then, of arriving at utilitarianism (although not, of course, the only way of doing so) is to adopt for society as a whole the principle of rational choice for one man.[1]

1.8.2 Rawls's idea here is that the idea of applying the principle of rational choice for one man occupies an analogous position in the construction of utilitarianism as the device of the hypothetical contract does in the construction of the two principles.[2] It is arguably rational for an individual to accept lower prospects of life than they had to for some periods of their life in order to have greater or more prolonged enjoyment at other stages. So it could be rational for someone to work extremely hard and not have much fun for a few years in order to qualify as a doctor and enjoy a comfortable life thereafter. Or it could be rational to party now and suffer later. Here is an illustration of this second alternative with a somewhat fanciful example. Suppose you suffered from a genetic disorder that you knew would strike you sometime in the last few years of your life, but you don't know when it will strike. This genetic disorder would make your final years 'not worth living' (though not absolutely unbearable) unless you made major sacrifices throughout your life to provide for those final years, lifting them over the 'worth living' threshold. Imagine also that this condition when it struck would make you paralyzed so suicide wasn't an option. Then it is arguably rational to choose the alternative of having a pretty good life for most of your life

[1] Rawls *T of J Rev* 1999 pp. 23-24

[2] In 'Distributive Justice: Some Addenda' (1968), Rawls writes 'the idea of the initial contractual situation involves many elements and can be defined in various ways. This situation is the analogue, in the contract theory, of the point of view of the impartial sympathetic spectator in utilitarianism' (p.69). Since, as I recount below, Rawls, in *Theory*, maintains that the device of the impartial sympathetic spectator amounts to the same thing as the device of adopting the principle of rational choice for one man, my description of the device of applying the principle of rational choice for one man as occupying an analogous position to the device of the hypothetical contract seems fair.

with a below-worth-living stretch for a few years at the end rather than having a mediocre but just-worth-living life throughout your life.

1.8.3 Although, as **Passage 1g *(T of J Rev)*** shows, Rawls concedes that utilitarianism could be arrived at from other starting points, he generally conceives of it in *Theory* as being founded on the idea of adopting for society the principle of rational choice for one man, and maintains that it is on this basis that utilitarianism might claim to be the most rational conception of justice.[1] I shall refer to this starting point as **the utilitarian conception of society.**

1.8.4 To illustrate how unlikely it might seem that Justice as Benevolence and Justice as Reciprocity would recommend the same principles to govern the well-ordered society, we can consider what the consequence of applying the principle of rational choice for one man might turn out to be. As the example in §1.8.2 illustrated, it might be rational for an individual to be prepared to live a proportion of their life – perhaps 10% - at a not-worth-living level for the sake of living a pretty good life for a much larger proportion of their life. But applying the same principle across society might make it rational for society to allocate resources so that 10% of the population had no prospect of having lives worth living while a much larger percentage enjoyed the prospect of leading pretty good lives! It is hard to see how a motive of reciprocity could apply to this 10%. If this 10% were to be motivated to act justly at all it could only be through sympathy, for thought of the damage their failure to act justly would do to the better off, rather than reciprocity. (And it would be hard for society to cultivate the kind of sympathy required to motivate that 10% to act justly.) This example seems to me to provide a fairly strong counterexample to both the utilitarian theory of justice and to utilitarianism's potential to be reconciled with Justice as Reciprocity. It provides a counter example to the utilitarian theory of justice because it seems intuitively unjust. So if it is the result of applying the utilitarian theory of justice, this provides an argument against that conception of justice and in favour of an alternative conception of justice such as Justice as Reciprocity. It also argues against the possibility of

[1] In fact, he only adopted this starting point for utilitarianism with his 1963 essay 'Constitutional Liberty and the Concept of Justice.' It plays no part in his two versions of 'Justice as Fairness.' The changing nature of Rawls's conception of utilitarianism will receive more attention in Chapter 2.

reconciling utilitarianism with Justice as Reciprocity, since it is hard to see how those 10% could have any motive to comply with rules of justice in repayment for the advantages they received from the compliance of others – 'repayment for what?' they might ask.

1.8.5 Although Rawls commonly conceived of utilitarianism as being derived from the utilitarian conception of society, he also contemplated its having its origin with the classical utilitarian idea of the 'impartial spectator'. In fact, he argued that the two starting points amounted to the same thing. So, immediately continuing on from **Passage 1g *(T of J Rev)***[1], Rawls wrote

The impartial spectator

Passage 1h *(Tof J Rev)*

> Once this is recognized, the place of the impartial spectator and the emphasis on sympathy in the history of utilitarian thought is readily understood. For it is by the conception of the impartial spectator and the use of sympathetic identification in guiding our imagination that the principle for one man is applied to society. It is this spectator who is conceived as carrying out the required organization of the desires of all persons into one coherent system of desire; it is by this construction that many persons are fused into one. Endowed with ideal powers of sympathy and imagination, the impartial spectator is the perfectly rational individual who identifies with and experiences the desires of others as if these desires were his own. In this way he ascertains the intensity of those desires and assigns them their appropriate weight in the one system of desire the satisfaction of which the ideal legislator then tries to maximize by adjusting the rules of the social system.[2]

1.8.6 I am not persuaded that the device of the impartial spectator and what I am referring to as the utilitarian conception of society are as intimately connected as Rawls implies. But resolving this issue is not important for my argument in this thesis. What is important to note is that Rawls appears to conceive of either of these as a foundation for what I shall refer to as **the utilitarian theory of justice.** This is personified by the figure of **the ideal legislator**, who adjusts the rules of the social system so as to maximize utility. The 'ideal

[1] p. 46
[2] Rawls *T of J Rev* 1999 p. 24

legislator' will play an important role in my arguments of Chapter 4. The utilitarian theory of justice that he represents, according to Rawls

Passage 1i *(T of J Rev)*

> receives perhaps its clearest and most accessible formulation in Sidgwick. The main idea is that society is rightly ordered, and therefore just, when its major institutions are arranged so as to achieve the greatest net balance of satisfaction summed over all the individuals belonging to it.[1]

1.8.7 This can usefully be viewed as Rawls's canonical statement of **the utilitarian theory of justice**. According to the utilitarian theory of justice then, acting justly would presumably be to act in accordance with the rules of institutions arranged to achieve the greatest net balance of satisfaction summed over all the individuals belonging to society.

1.8.8 The utilitarian theory of justice provides a useful intermediary device for the argument of my thesis; particularly for the argument of Chapter 4 when I consider Rawls's argument that utilitarianism is unjust because it fails to take seriously the 'separateness of persons'. It is important, I believe, to distinguish at exactly what component of Justice as Benevolence that charge is levelled at. The charge that the utilitarian conception of justice fails to take the separateness of persons seriously might turn out to be different to the charge that the utilitarian theory of justice fails to take the separateness of persons seriously. In Chapter 4, I shall maintain that it is.

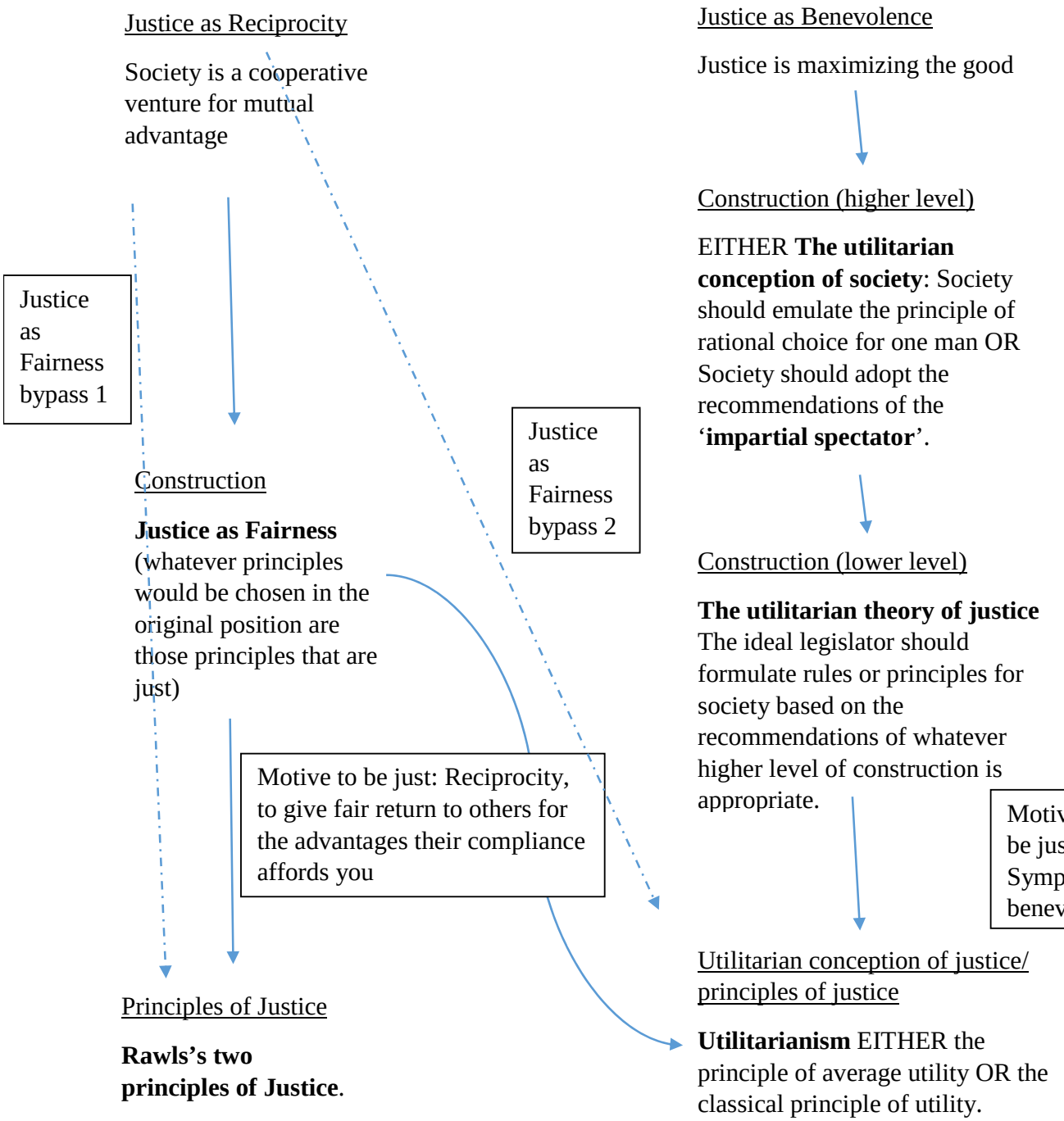
1.8.9 In this section and the last I have fleshed out the details of Rawls's views of the two conceptions of justice, Justice as Benevolence and Justice as Reciprocity, which provided the focus of the passages in the Preface. These accounts run from the level of construction to the principles that are supposed to govern the well-ordered society of either conception. In response to the question of what would motivate citizens in the well-ordered society governed by the two principles of justice to act justly, Rawls's answer was 'reciprocity'. In response to the question of what would motivate citizens to act justly in the well-ordered society governed by the principle of utility. Rawls's answer was 'sympathy'.

[1] Rawls *T of J Rev* 1999 p. 20

1.8.10 So it is easy to see how, given Rawls's understanding of classical utilitarianism as starting from either the utilitarian conception of justice or the impartial spectator and his own starting point of justice being a matter of reciprocity, it must have seemed highly unlikely that utilitarianism would be reconcilable with reciprocity.

1.8.11 I end this section by providing a diagrammatic summary of Rawls's conceptions as described so far in this Chapter with two significant additions: the first is **bypass 1**, which goes straight from the conception of society as a cooperative venture for mutual advantage to the two principles of justice. This is an alternative route that Rawls's first model of his theory might have taken as I explain in Chapter 2. In Chapter 4 I suggest that **bypass 2** may lead from Justice as Reciprocity to the utilitarian conception of justice.

Fig 1(i)


Justice as Reciprocity
Society is a cooperative venture for mutual advantage
Justice as Benevolence
Justice is maximizing the good
Construction (higher level)
EITHER The utilitarian conception of society: Society should emulate the principle of rational choice for one man OR Society should adopt the recommendations of the 'impartial spectator'.
Justice as Fairness bypass 1
Justice as Fairness bypass 2
Construction
Justice as Fairness (whatever principles would be chosen in the original position are those principles that are just)
Construction (lower level)
The utilitarian theory of justice The ideal legislator should formulate rules or principles for society based on the recommendations of whatever higher level of construction is appropriate.
Motive to be just: Reciprocity, to give fair return to others for the advantages their compliance affords you
Utilitarian conception of justice/ principles of justice
Principles of Justice
Rawls's two principles of Justice.
Utilitarianism EITHER the principle of average utility OR the classical principle of utility.

9 The Problems

1.9.1 In my Preface I described how when I first read *Theory* I couldn't make sense of certain passages because they *didn't make sense*, and that their lack of sense was due to their being the result of attempts to patch up arguments that Rawls had put forward in previous essays, by substituting new premises for old; the old premises still lying, largely neglected but still intact, in those earlier essays. I also described how my method in this thesis would be to expose the problems by comparing the passages from *Theory* side by side with their historical antecedents. In this section I provide a taste of that method by comparing a passage from *Theory* side by side with its antecedent from 'Distributive Justice' (1967) in order to expose some of the problems Rawls was never able to solve. The comparison of the two passages reveals that Rawls's biggest problem was that he could never find a satisfactory alternative to defining mutual advantage in comparison to the state of nature outlined by Thomas Hobbes in *Leviathan*. I should state at this point that, following Wolff, I shall refer to Rawls's theory as it was in 'Distributive Justice' as the 'second model' of his theory, and as it was in *Theory*, as the third model.

1.9.2 In the opening few pages of *Theory*, Rawls put a brief case for the rejection of the principle of utility that he clearly intended to be intuitively appealing to the reader in advance of his more detailed argument. This is repeated below

Passage 1j *(T of J Rev)* The argument of *Theory*

> [A] It may be observed, however, that once the principles of justice are thought of as arising from an original agreement in a situation of equality, it is an open question whether the principle of utility would be acknowledged. [B] Offhand, it hardly seems likely that persons who view themselves as equals, entitled to press their claims upon one another, *would agree to a principle that may require lesser life prospects for some simply for the sake of a greater sum of advantages enjoyed by others.* [C] Since each desires to protect his interests, his capacity to advance his conception of the good, no one has a reason to acquiesce in an enduring loss for himself in order to bring about a greater net balance of satisfaction. [D] In the absence of strong and lasting benevolent impulses, a rational man would not accept a basic structure merely because it maximized the algebraic sum of advantages irrespective of its permanent effects on his own basic rights and interests. [E] Thus it seems that the principle of utility is incompatible with the idea of

> *reciprocity* implicit in the notion of a well-ordered society. [my letters and italics][1]

1.9.3 When a writer introduces an argument with a rhetorical flourish to the effect that it should be obvious, as Rawls did in Sentence [B], that is often a signal that it will turn out to be anything but obvious. So it proves in this case, as I show below.

1.9.4 To set the argument of **Passage 1j *(T of J Rev)*** in context, Rawls had already outlined his theory of Justice as Fairness so the reader first approaching this passage would be aware that the original agreement referred to in [A] was between parties situated behind a veil of ignorance that obscured their knowledge of the social positions or conceptions of the good of the real people they represented.[2] So the first question that should strike the reader regarding the italicized claim in [B] is 'why wouldn't they agree to a principle that would requires lesser life prospects for some for the sake of a greater sum of advantages enjoyed by others?' They don't know whether they represent the 'some' who will turn out to have 'lesser life prospects' or the 'others' who enjoy the greater sum of advantages. It will be easier to address this question and others arising from the passage after looking at the argument of the equivalent passage from Rawls's earlier essay.

Passage 1k (*DJ 1967)* The argument of 'Distributive Justice'

> [1] Once justice is thought of as arising from an original agreement of this kind, it is evident that the principle of utility is problematical. [2] For why should rational individuals who have a system of ends they wish to advance agree to *a violation of their liberty for the sake of a greater balance of satisfactions enjoyed by others?* [3] It seems more plausible to suppose that, when situated in an original position of equal right, they would insist upon institutions which returned compensating advantages for any sacrifices required. [4] A rational man would not accept an institution merely because it maximized the sum of advantages irrespective of its effect on his own interests. [My italics and numbering of the sentences] [3]

1.9.4 It should be fairly obvious that this passage can be viewed as a historical antecedent

[1] Rawls *T of J Rev* 1999 p.13
[2] Rawls *T of J Rev* 1999 pp.10 - 12
[3] Rawls DJ 1967 p.132

of **Passage 1j *(T of J Rev).*** Its claim in Sentence [1] that the principle of utility is 'problematical' is similar to the first clause of Sentence [B]'s claim that the acknowledgement of the principle of utility is 'hardly likely'. And the justification of [1]'s claim that the principle of utility is problematical in the italicized part of [2] could be read as occupying the same position in the argument of **Passage 1k (*DJ 1967)*** as the italicized second clause of [B] in **Passage 1j *(T of J Rev).***[1] The main conclusion of the two passages is essentially the same. The implicit conclusion of [3] in **Passage 1k (*DJ 1967)*** is that the parties in the original position would reject the principle of utility in favour of other principles. The implicit conclusion of [E] in **Passage 1j *(T of J Rev)*** is that the parties in the original position would reject the principle of utility in favour of other principles.[2]

1.9.5 But now it can be seen that the two passages, while supporting the same conclusion, offer quite different premises in support of the main conclusion. The argument of *Theory* concludes that the parties in the original position would reject the principle of utility because they wouldn't accept a principle that *that may require lesser life prospects for some* for the sake of *a greater sum of advantages enjoyed by others.* The argument of 'Distributive Justice' concludes that the parties in the original position would reject the principle of utility because they wouldn't accept a principle that would *violate their liberty* without giving them *compensating advantages* in return.

1.9.6 My interpretation of the argument of **Passage 1k (*DJ 1967)*** calls for some more justification. Rawls is suggesting that it is highly unlikely that the parties in the original position would choose the principle of utility, and part of the support for that conclusion is that rational individuals who were concerned with their own ends would not agree to *a violation of their liberty for the sake of a greater balance of satisfactions enjoyed by others.* Now, one might be inclined to read Rawls as asserting here that the parties in the original position would not agree to a violation of their liberty for any reason, not just for the sake

[1] p. 52

[2] This conclusion is not quite so clear in the passage from *Theory* as it is in the passage from 'Distributive Justice'. This is because Sentence [3] from **DJ 1967** gives the parties decision as the reason for the rejection of the principle of utility in the same sentence, which Sentence [E] from *T of J Rev* doesn't. But sentence [E] is obviously intended to follow as a conclusion from the rest of the passage, all of which concerns the decision of the parties in the original position.

of a greater balance of satisfactions enjoyed by others. This reading might construe Rawls to be asserting that people were already in possession of inviolable rights that the principle of utility would be liable to violate.[1]

1.9.7 But the following line, [3], doesn't fit with such an interpretation. [3] doesn't say that the parties in the original position would not agree to have their liberty violated for any reason. Instead, it appears to be asserting that the parties in the original position *would* agree to a violation of their liberty, *if* they received compensating advantages for that violation of their liberty. The parties' agreed-to 'violation of their liberty' is what Rawls is naturally read to mean by [3]'s 'the sacrifices required'. This second interpretation would bring Rawls in line with the traditional social contract theorists, Hobbes, Locke and Rousseau, who all held the social contract to require people to surrender some, if not all, of their natural liberty.[2]

1.9.8 The second interpretation of Rawls's argument which, henceforth, I shall assume to be correct, carries with it two important implications. The first implication is that the principle of utility would not compensate the parties for their loss of liberty, and the second is that Rawls's principles of justice would, by contrast, provide such compensation. The

[1] In fact, as I describe in Chapter 4, Rawls did briefly contemplate an original position in which the parties involved were already in possession of their rights which the principle of utility would be liable to violate. That was in his essay 'Constitutional Liberty and the Concept of Justice' (1963) and I refer to the state Rawls's theory was in at that time as model 1.5. That is because it lies halfway between what Wolff in *Understanding Rawls* describes as Rawls's first model of 'Justice as Fairness' and his second model of 'Distributive Justice'. (Wolff 1977)

[2] Locke certainly held that the social contract required people to give up their liberty to punish transgressors of the law of nature, while retaining their liberty to acquire and dispose of private property, so should be interpreted as a social contract theorist who required people to surrender some, but not all, of their natural liberty. (See Locke 2003 *Two Treatises of Government* Chapters 8&9) Rousseau held that the social contract required people to give up all of their 'natural liberty' in exchange for 'civil liberty', so should be regarded as viewing the social contract as requiring compete surrender of their natural liberty. (See Rousseau 2002 *The Social Contract* Book 1 Chapter 6). Hobbes is more complicated and ambiguous. He claimed that it would be rational for people to promise away all their natural liberty, but then could be viewed as holding that they retained some of their natural liberty nonetheless, as he held it to be rational for rebels to fight for survival against a ruler intent on their death (See Hobbes 1996 *Leviathan* Part 2 Chapter 21). Hobbes' position is particularly pertinent to my thesis as I shall argue in Chapter 4 that the underlying logic of justice as reciprocity should be committed to a similar position, that those who are not afforded a sufficient prospect of a life worth living through social co-operation, have no obligation to cooperate with society.

second of these is implied by [3]'s implication that the parties would have the option of insisting on institutions which did return compensating advantages for their loss of liberty.

1.9.9 We can put aside the question of what 'the liberty of the original position', as I shall refer to the liberty invoked in **Passage 1k *(DJ 1967)***[1], would amount to for now, in order to question whether Rawls's conclusion of sentence [2] is as obvious as he implies it is.

1.9.10 And it turns out not to be, for reasons related to the design of the original position. This design was essentially the same in 'Distributive Justice' as it was in *Theory*. As Rawls described it then

Passage 1l: The original position in 'Distributive Justice' *(DJ 1967)*

> the contract doctrine assumes that the rational individuals who belong to society must choose together, in one joint act, what is to count among them as just and unjust. They are to decide among themselves once and for all what is to be their conception of justice. This decision is thought of as being made in a suitably defined initial situation one of the significant features of which is that no one knows his position in society, nor even his place in the distribution of natural talents and abilities. The principles of justice to which all are forever bound are chosen in the absence of this sort of specific information. A veil of ignorance prevents anyone from being advantaged or disadvantaged by the contingencies of social class and fortune; and hence the bargaining problems which arise in everyday life from the possession of this knowledge do not affect the choice of principles.[2]

1.9.11 As I interpret **Passage 1k (*DJ 1967)*** the argument is that it is unlikely that the parties would choose the principle of utility not because *all of them* would fail to be compensated adequately for their loss of liberty but because *some of them* would. This reading interprets **Passage 1k (*DJ 1967)*** as assuming that there is a fixed sum of people, and that that 'the principle of utility' is the principle of average utility, which is reasonable given Rawls's later justification of these assumptions dealing with the same scenario in *Theory*. Given these assumptions, the losses to those people who end up worse off than they would be in a situation of equal liberty translate into gains for others who do better

[1] p. 53
[2] Rawls DJ 1967 p. 132

than they would in a situation of equal liberty. But, as Rawls says, one of the 'significant features' of the original position is that 'no one knows his position in society'. In that case, if we assume that the parties really don't know what position they would occupy in society, those who would fail to be adequately compensated for their loss of liberty - call them 'the losers' - by the principle of utility would not know themselves to be such. They might instead turn out to be one of those others who might do better under the principle of utility than under Rawls's principles of justice – call them 'the winners'.

1.9.12 In the light of their lack of knowledge of whether they would turn out to be a winner or a loser, a rational individual *might* choose the principle of utility since this is the principle that could be expected to work out best for them. They would be prepared to run the risk of turning out to be amongst those who do not receive adequate compensation for their loss of liberty (i.e. a loser) for the prospect of greater gains to those who do (i.e. a winner). Whether or not that would actually be the rational choice is debatable, but I don't need to enter that debate in order to point out that the implication of Rawls's argument, that it would be unlikely that people in the original position would choose the principle of utility because it would fail to compensate *some* of them for violation of their liberty, is highly questionable.

1.9.13 How should the fact that that it may be rational for the parties in the original position to choose the principle of utility affect the rest of the claims of this passage? Apart from the point already made, that they might choose the principle of utility over the two principles of justice, there are three further remarks to be made. Let us assume, for the sake of argument, that persons in the original position would indeed choose the principle of utility. The first remark to be made, then, is that it would be inaccurate to describe the parties in the original position as rational individuals who are agreeing to *a violation of their liberty for the sake of a greater balance of satisfactions enjoyed by others* as [2] does. It would be more accurate to describe them as agreeing to run the risk of losing their liberty for *the expectation of greater gains to themselves.* The second remark is that the claim of [3] would be similarly misleading. The parties would be prepared to run the risk of institutions which did not return compensating advantages for their sacrifice of liberty. The third remark concerns the accuracy of sentence [4]. The implication is that all the principle

of utility would do is maximize the sum of advantages, and the parties in the original position's interests would be better served elsewhere, either by the liberty of the original position or alternative principles of justice. But if the parties in the original position chose the principle of utility because they felt the gamble was rationally justified they should not be described as doing so irrespective of its effect on their interests. Rather, they should be described as accepting utilitarian institutions, *in the hope that doing so would further the pursuit of their own interests.*

1.9.14 So it is by no means clear that the parties in the original position of 'Distributive Justice' would reject the principle of utility. But it is worth seeing what can be salvaged from the argument of **Passage 1k (*DJ 1967*)**[1] and **Passage 1l *(DJ 1967)***[2], and whether what can be salvaged would be enough to condemn the principle of utility from the perspective of Justice as Reciprocity. The purpose of this salvaging operation is firstly; to provide a coherent argument which condemns the principle of utility from the perspective of Justice as Reciprocity that can be used as point of contrast for the incoherent arguments of *Theory*, examined below (§§1.9.44 –1.9.52), and secondly; to expose two problems for Rawls's theory of Justice as Reciprocity that I maintain he was unable to resolve.

1.9.15 I shall introduce this salvaging operation by observing that the difficulty Rawls's argument had in establishing its conclusion was that it was arguably rational for people in the original position to choose a principle that might fail to compensate some cooperating members of society for their loss of liberty. But this difficulty seems to point to a problem with the design of the original position for its intended purpose of coming up with the principles best suited for the conception of Justice as Reciprocity. It might be argued that Justice as Reciprocity should surely require that *every* individual cooperating member of society receive adequate compensation for their loss of liberty. This line of argument points to an obvious flaw in the design of the original position; it is the veil of ignorance which appears to be flawed.

1.9.16 To appreciate how the veil of ignorance might undermine the point of the original

[1] p. 53
[2] p. 56

position, it is helpful to recall what the ultimate purpose of the original position is. This is to construct principles suited for the conception of society as a cooperative venture for mutual advantage. As **Passage 1l *(DJ 1967)*** above reports, Rawls's stated aim for the veil of ignorance was to prevent 'the bargaining problems which arise in everyday life from the possession of this knowledge [of one's position in society and place in the distribution of natural talents and abilities]' from affecting the choice of principles. This aim easily fits with Justice as Reciprocity. If people know too much about each other no agreement could be reached, and a cooperative venture for mutual advantage might seem impossible to achieve. But the veil of ignorance appears to overshoot its aim by preventing those who wouldn't receive adequate compensation for their loss of liberty under particular principles from knowing this and being able to veto principles in light of that knowledge.

1.9.17 Thomas Nagel neatly pointed to the fundamental problem with Rawls's veil of ignorance, back in 1971 in *The Possibility of Altruism.* Although his concern was with the viability of the hypothetical contract for the purpose of constructing a reasonable theory of altruism rather than of Justice as Reciprocity, the problem he pointed out also has implications the viability of Justice as Fairness as an appropriate device for constructing principles of reciprocity. Nagel observed

Passage 1m (*P of A)*: Nagel's criticism of the original position

> it will be natural for the person [i.e. party in the original position] choosing to think of the various lives, one of which he is already settled with, as *possibilities;* it is possible that he is a slave, but then again it is possible that he is a master. And he may be able to tolerate as an outcome of the interpersonal weighting system a small percentage of heavy losers. Such tolerance seems to deny the interests of these people due weight, since there really *are* individuals in these roles, and their lives are not possibilities, but actualities: the only lives they have.[1]

1.9.18 Nagel's argument can be adapted to the context of Justice as Reciprocity as follows. As Nagel says, the device of choice from behind a veil of ignorance is liable to treat actual people as possibilities. The party in the original position might take a gamble on becoming

[1] Nagel 1970 p.140

a slave or a master, but the result of his gamble would be to impose slavery on a real person. And that real person, the slave, might not receive compensating advantages for their loss of liberty.

1.9.19 As mentioned above (§1.9.15) it might seem a reasonable requirement of Justice as Reciprocity that all cooperating members of society receive compensation for their sacrifice of liberty, and to the extent that the veil of ignorance interferes with the original position's ability to ensure that requirement is met, it is a hindrance rather than a help.

The more promising reconstruction of Rawls's argument of 'Distributive Justice'.

1.9.20 In view of all the subsequent revisions of his theory, and the revision of this particular passage which Rawls was evidently unhappy with, it is reasonable to speculate that Rawls may not have fully appreciated the hindering aspect of the veil of ignorance when he wrote **Passage 1k (*DJ 1967)***[1], and that he had assumed that principles that failed to compensate all for their loss of liberty could not fetch agreement amongst the parties in the original position.[2] So let us suppose that Rawls had somehow found a way to circumnavigate the problem under consideration, and devise an original position with a veil of ignorance that prevented 'the winners' (those who would receive compensating advantages for their loss of liberty) from knowing which principles would work out more to their advantage, given their particular circumstances, while 'the losers' (those who would

[1] p. 53

[2] There is plenty of circumstantial evidence to support this speculation. I have already indicated my agreement with Wolff's view that '[t]he labyrinthine complexities of *A Theory of Justice* are the consequences of at least three stages in the development of Rawls's thought, in each of which he complicated his theory to meet objections others had raised to earlier versions, or which he himself perceived.' 'Distributive Justice' is what Wolff takes to be the defining essay of Rawls's second model (Wolff 1977 p.5) where Rawls has just introduced the veil of ignorance. It follows from a model I have identified as model 1.5 (to be examined in detail in Chapter 4) in 'Constitutional Liberty and the Concept of Justice' (1963) where Rawls conceives of the people in the original position as being already in possession of their basic liberties, and having knowledge of their identities, requiring genuine unanimous agreement. Rawls, in my view, had not managed to properly 'move on' from his previous model. A third commentator, beside myself and Wolff, who has commentated on the inconsistency of Rawls's work is H.L.A. Hart who, in 'Rawls on Liberty and its Priority' (1973 p. 541), remarks on 'difficulties in this interpretation [of certain claims in *Theory*] which suggest that Rawls has not eliminated altogether the earlier general doctrine of liberty.' The doctrine referred to by Hart is the doctrine of Rawls's first model which receives examination in Chapter 2 of this thesis.

not receive compensating advantages for their loss of liberty) knew themselves to be such and were able to reject any principles that would lead them to lose. We can then salvage a more promising line of argument against the principle of utility from the perspective of Justice as Reciprocity. According to this new line of argument, the principle of utility would be rejected by the parties in the original position because it failed to ensure that each and every cooperating member of society received adequate compensation for their loss of liberty.

1.9.21 The argument, thus reconstructed, is better placed to uphold the main conclusion of **Passage 1k (*DJ 1967)*[1]**, that the parties in the original position would reject the principle of utility. An implicit claim of [2], that by agreeing to the principle of utility the losers would be agreeing to a sacrifice of their liberty for the sake of others could be upheld. They would be better off if all retained the liberty of the original position, so agreeing to the principle of utility would require that they sacrifice their liberty for the sake of the benefits that their loss of liberty would provide to 'others', namely, the winners. So the second implicit claim of [2], that the parties in the original position would not agree to the principle of utility because the losers would not agree to lose their liberty if they received no compensating advantages in return also seems plausible. The claim of [3] could also be upheld. Rational individuals who were concerned to advance their ends, and not to provide benefits to others, would insist on institutions (such as those governed by the two principles of justice) that would provide compensating advantages to themselves for their loss of liberty.

1.9.22 So in summary: the 'more promising' reconstruction of Rawls's argument of the last two paragraphs (§§1.9.20 - 1.9.21), which supposes that the 'losers' could and would veto principles under which they would lose, would appear to be both more suited to the conception of Justice as Reciprocity and better able to sustain the claims that the principle of utility is incompatible with the conception of Justice as Reciprocity than the interpretation of Rawls argument examined earlier (§§1.9.4 –1.9.13) in which the losers' knowledge that they were losers was hidden from them by the veil of ignorance.

[1] p. 53

1.9.23 However, there are two serious problems with Rawls's argument as thus reconstructed, one of which he was evidently aware of at the time, the second of which he evidently became aware of later.

The first problem

1.9.24 The first problem, of which Rawls was certainly aware, is that taking the baseline of 'the liberty of the original position' as the baseline from which the 'compensating advantages' provided by principles of distributive justice might be measured, might not condemn as 'unjust', institutions that we intuitively feel deserve that designation. So, in a highly revealing passage from 'Distributive Justice', that will receive more detailed examination in Chapter 3, Rawls wrote

Passage 1n (*DJ 1967*): Rawls on Hume's definition of mutual advantage

> But all Hume seems to mean by this [the possibility of defining 'advantage' in comparison to some historically relevant benchmark] is that everyone is better off in comparison with the situation of men in the state of nature, understood either as some primitive condition or as the circumstances which would obtain at any time if the existing institutions of justice were to break down. While this sense of everyone's being made better off is perhaps clear enough, Hume's interpretation is surely unsatisfactory. For even if all men including slaves are made better off by a system of slavery than they would be in the state of nature, it is not true that slavery makes everyone (even a slave) better off, at least not in a sense that makes the arrangement just. The benefits and burdens of social cooperation are unjustly distributed even if everyone does gain in comparison with the state of nature; this historical or hypothetical benchmark is simply irrelevant to the question of justice. In fact, any past state of society other than a recent one seems irrelevant offhand, and this suggests that we should look for an interpretation independent of historical comparisons altogether. Our problem is to identify the correct hypothetical comparisons defined by currently feasible changes.[1]

[1] Rawls DJ 1967 p. 135

1.9.25 The point for now is that, depending on how the 'liberty of the original position' is to be defined, it might allow slaves to qualify as being 'better off' under a system of slavery than in a state in which all enjoyed the liberty of the original position and that, according to Rawls, would not do since slavery is unjust.[1] So the first problem with taking the baseline of a state of nature to be the baseline by which to measure mutual advantage is that it might allow slavery to qualify as just.

1.9.26 This point can be confirmed by my delineating a Hobbesian definition of natural liberty and the state of nature that it would seem to define in more detail. This will also serve the purpose of casting light on the other problem raised by the reconstructed argument of **Passage 1k *(DJ 1967)*[2].**

1.9.27 Hobbes defined the 'Right of Nature' in *Leviathan* as 'the Liberty each man hath, to use his own power, as he will himselfe, for the preservation of his own Nature; that is to say, of his own Life; and consequently, of doing any thing, which in his own Judgement, and Reason, hee shall conceive to be the aptest means thereunto.'[3] This definition confuses the issue a bit from the perspective of my attempt to construct a definition of natural liberty suitable for the conception of Justice as Reciprocity, as it might be taken to imply that people didn't have the natural right to do anything to advance ends they might have that aren't simply the preservation of their life. In light of this, I shall modify Hobbes's actual definition to come up with a 'Hobbesian' definition of natural right that simply allows one to use one's power to do anything to advance any purpose whatsoever. Just how broadly Hobbes intended this right to do 'any thing' to be understood, becomes apparent shortly after he offered this definition when he goes on to generalize the implications of this right for his State of Nature, which, as he had explained earlier in *Leviathan*, is a war of all against all.

Passage 1o (*Lev*)

[1] In Chapter 3 I argue that Rawls's first argument for his difference principle should be understood as an attempt to get round this problem.
[2] p. 53
[3] Hobbes 1996 p.91

> And because the condition of Man, (as hath been declared in the previous Chapter) is a condition of Warre of every one against every one; in which case every one is governed by his own Reason; and there is nothing he can make use of, that may not be a help unto him, in preserving his life against his enemyes; It followeth, that in such a condition, every man has a Right to every thing: even to one anothers body. [1]

1.9.28 For my modified Hobbesian definition we should ignore the clause that might seem to limit the use of one's natural right to preserve one's life against one's enemies and focus on the fact that it includes the freedom to use one another's body which, by anyone's standard, is a very permissive interpretation of natural liberty. It is also in stark contrast to the natural liberty that John Locke supposed we were once entitled to, whereby

Passage 1p (*Two Treatises)*

> a state of liberty is not a state of licence: though man in that state have an uncontrolleable liberty to dispose of his person or possessions, yet he has not liberty to destroy himself, or so much as any creature in his possession, but where some nobler use than its bare preservation calls for it. The state of nature has a law of nature to govern it, which obliges every one: and reason, which is that law, teaches all mankind, who will but consult it, that being all equal and independent, no one ought to harm another in his life, health, liberty, or possessions:[2]

1.9.29 In one way, the original position described in **Passage 1k *(DJ 1967)*** [3] and **Passage 1l** (***DJ 1967)*** [4] is more reminiscent of Locke's state of nature than Hobbes's, as Rawls implies in **Passage 1k *(DJ 1967)***, that the parties would be able to advance their ends, at least to some extent, if they rejected any contract in favour of retaining the liberty of the original position. But in another way it is more reminiscent of Hobbes's social contract as Locke, in **Passage 1p (*Two Treatises)***, maintains that people in the state of nature have 'possessions', in other words 'property', which is protected by the law of nature. For Rawls, as for Hobbes, there is no protection of property prior to the social contract. This thesis shall maintain that Rawls's original position is rather more like Hobbes's state of

[1] Hobbes 1996 p.92
[2] Locke 2003 p.102
[3] p. 53
[4] p. 56

nature than Rawls would like, as there is no rationale within his theory for *any* restraints on natural liberty prior to the social contract.

1.9.30 Hobbes's state of nature was, however, in two important ways very different to Rawls's original position. First, as already remarked above (§1.9.27), Hobbes supposed people to have no other end than their self-preservation. Rawls, by contrast, is open to their having a variety of ends in **Passage 1k *(DJ 1967)*[1]**. Secondly, Hobbes took it to be simply a matter of fact that people in a state of nature would be roughly equal in power so no one individual could expect to dominate another. Rawls supposed 'free and independent persons in an original position of equality' in order to 'reflect the integrity and equal sovereignty of the rational persons who are the contractees'.[2] The results of Rawls's hypothetical contract are intended to be applied to people who as a matter of fact might well not be free, independent, equal or even rational. There is an important moral premise underlying Justice as Fairness (at least on my, and many other interpreters of Rawls); the contract treats people as free and equal because this is how they *ought* to be treated. There is no such moral premise underlying Hobbes's social contract theory; indeed, Hobbes maintains that notions of morality have no place in a state of nature.[3]

1.9.31 But important though these differences may be between Hobbes' and Rawls's social contract theories are, they need not prescribe different definitions of natural liberty. There is nothing in the Hobbesian definition of liberty I have given so far that contradicts the aims of Justice as Fairness. Assuming that the parties in the original position would all have liberty to do whatever they wanted in pursuit of their various conceptions of the good is, at least arguably, to reflect the integrity and equal sovereignty of the rational persons who are the contractees, regardless of the fact that Hobbes did not produce his definition of the right of nature with Justice as Fairness in mind. In fact, it is the definition of equal natural liberty that Rawls later appears to be committed to in *Theory*

Passage 1q (*T of J Rev)*: no-agreement point as baseline

[1] p. 53
[2] Rawls DJ 1967 pp. 131 - 132
[3] Hobbes 1996 p. 90

> To be sure, from the standpoint of the original position, the principles of justice are collectively rational: everyone may expect to improve his situation if all comply with these principles, at least in comparison with what his prospects would be in the absence of any agreement. General egoism represents this no agreement point.[1]

1.9.32 This no agreement point of general egoism is *Theory*'s equivalent of what I have referred to as 'the liberty of the original position' of 'Distributive Justice'.

1.9.33 What would a state of nature of general egoism look like? It would be a shame to waste the opportunity to describe it in the words of the political philosopher who did most to warn the world to avoid such a possibility, so I won't. As Thomas Hobbes put it in Leviathan

Passage 1r (*Lev*): Hobbes's state of nature

> Whatsoever therefore is consequent to a time of Warre, where every man is Enemy to every man; the same is consequent to the time, wherein men live without other security, than what their own strength, and their own invention shall furnish them withall. In such condition, there is no place for Industry; because the fruit thereof is uncertain: and consequently no Culture of the Earth; no Navigation, nor use of the commodities that may be imported by Sea; no commodious Building; no Instruments of moving, and removing such things as require much force; no Knowledge of the face of the Earth; no account of Time; no Arts; no Letters; no Society; and which is worst of all, continuall feare, and danger of violent death; And the life of man, solitary, poore, nasty, brutish and short.[2]

Second problem

1.9.34 The preceding paragraphs should make it easier to describe the second of the problems that the reconstruction of Rawls's argument of 'Distributive Justice' would have faced. As Hobbes pointed out, a state of general egoism would be pretty dire, dire enough to make it *probable* that the principle of utility *would* improve *everyone's* prospects by

[1] Rawls *T of J Rev* 1999 p. 435
[2] Hobbes 1996 p. 89

comparison. In which case, the 'more promising' argument considered earlier ((§§1.9.20 - 1.9.21) that the parties in the original position would reject the principle of utility because it failed to compensate all for their loss of the liberty of the original position would not hold water.

1.9.35 This problem Rawls certainly appeared to be aware of by the time of writing *Theory,* for there he is explicitly committed to the position that the principle of utility *would* improve the prospects of all in comparison with the benchmark of the state of nature. So he wrote

Passage 1s (*T of J Rev)*: all conceptions of justice superior to general egoism

> …it is obvious that by choosing one of the other conceptions the persons in the original position can do much better for themselves. Once they ask which principles all should agree to, no form of egoism is a serious candidate for consideration in any case.[1]

1.9.36 The context of this passage is that the parties in the original position (as it was in *Theory)* are comparing the alternative conceptions of justice available to them. Those alternatives include, amongst others, general egoism, the classical principle of utility, the principle of average utility and Rawls's two principles of justice.[2] So Rawls's statement that 'it is obvious that by choosing one of the other conceptions the persons in the original position can do much better for themselves,' should be taken to imply that by choosing either classical utilitarianism or the principle of average utility the parties could do better than under the conception of general egoism.

1.9.37 Therefore, by the same logic as Rawls held that, 'from the standpoint of the original position, the principles of justice are collectively rational'[3] because everyone may expect to improve his situation if all comply with these principles, at least in comparison with what his prospects would be in the absence of any agreement', so consistency should have

[1] Rawls *T of J Rev* 1999 p. 117
[2] Rawls *T of J Rev* 1999 p. 117
[3] Already cited in Passage 1q (*T of J Rev)* above.

obliged him to hold that it would have been collectively rational from the standpoint of the original position for the parties to choose the principle of classical utility or the principle of average utility. This is *because either of those conceptions of justice offered them a better alternative than the liberty of the original position.*

1.9.38 The position Rawls appears to be committed to in *Theory*, then, is in stark contrast to the claim of the reconstructed argument from 'Distributive Justice' which was that rational persons in the original position would reject utilitarianism because it failed to compensate them for their sacrifice of the liberty of the original position.

1.9.39 Here I recap the second problem posed for the reconstructed argument of 'Distributive Justice'. That argument looked promising because (as suggested in §1.9.15) it seems, at least at first sight, to be a reasonable requirement that principles of distributive justice should at least compensate all cooperating members of society for their loss of natural liberty. And, the reconstructed argument of 'Distributive Justice' disqualified the principle of utility from the competition on the grounds that it didn't compensate all cooperating members of society for their loss of natural liberty. But, according to Rawls's commitments of *Theory*, either utilitarian conception of justice *would* compensate all cooperating members of society for their loss of natural liberty. So they should not, after all, be disqualified as contenders for the principles of justice most suited to the conception of Justice as Reciprocity.

1.9.40 I turn now to re-examine the revision of the argument from 'Distributive Justice' that Rawls puts early in *Theory*. I shall argue that if the reconstructed argument of 'Distributive Justice' had serious problems, it was still far more coherent than anything that can be put together from the revised argument of *Theory*. For the reader's convenience, I repeat **Passage 1j *(T of J Rev)*** from the beginning of this section below

Passage 1j *(T of J Rev)*: The argument of *Theory*

> [A] It may be observed, however, that once the principles of justice are thought of as arising from an original agreement in a situation of equality, it is an open question whether the principle of utility would be acknowledged. [B] Offhand, it

> hardly seems likely that persons who view themselves as equals, entitled to press their claims upon one another, *would agree to a principle that may require lesser life prospects for some simply for the sake of a greater sum of advantages enjoyed by others.* [C] Since each desires to protect his interests, his capacity to advance his conception of the good, no one has a reason to acquiesce in an enduring loss for himself in order to bring about a greater net balance of satisfaction. [D] In the absence of strong and lasting benevolent impulses, a rational man would not accept a basic structure merely because it maximized the algebraic sum of advantages irrespective of its permanent effects on his own basic rights and interests. [E] Thus it seems that the principle of utility is incompatible with the idea of *reciprocity* implicit in the notion of a well-ordered society.' [my letters and italics][1]

First interpretation of the argument of *Theory*

1.9.41 The most significant difference to the equivalent argument of 'Distributive Justice', already remarked upon above (§1.9.5), is the italicized part of line B, where Rawls has substituted the principle of utility's property that it '*may require lesser life prospects for some simply for the sake of a greater sum of advantages enjoyed by others*' for the charge of 'Distributive Justice', that it may require '*a violation of their liberty for the sake of a greater balance of satisfactions enjoyed by others'*. The idea that principles appropriate to the conception of Justice as Reciprocity must avoid requiring lesser life prospects for some simply for the sake of a greater balance of satisfaction enjoyed by others is less reasonable than the idea that such principles should compensate all for their loss of natural liberty. I show this below.

1.9.42 An equivalent argument can be applied to the first interpretation of the argument from *Theory* as was applied to the first interpretation of the argument of 'Distributive Justice' in paragraphs (§§1.9.4 – 1.9.13). The parties in the original position would not know whether they would be the ones who would have lesser life prospects for some simply for the sake of others, as that information would be hidden from them by the veil of ignorance. Analogous reasoning applies here as to the reasoning of that argument of 'Distributive Justice' but according to the assumptions of *Theory*, the parties run the risk of being 'losers' who have to endure lower life prospects for the sake of others instead of being losers who are not adequately compensated for their loss of natural liberty. Since the

[1]Rawls *T of J Rev* 1999 p.13

reasoning is analogous to sections (§1.9.4 – §1.9.13), I do not need to go into it in such detail here. It is, however, worth remarking that Rawls has, in *Theory*, clarified the choice facing the parties as one between the two principles of justice and the principle of *average* utility, so the parties should have more confidence that their expected prospects would be greater under the principle of utility than under the two principles of justice.

1.9.43 So the assertion of sentence [B] that '[o]ffhand, it hardly seems likely that persons who view themselves as equals, entitled to press their claims upon one another, *would agree to a principle that may require lesser life prospects for some simply for the sake of a greater sum of advantages enjoyed by others'* provides, to my mind, a fine illustration of the point I raised in the Preface of Rawls concealing weak argument behind strong rhetoric. This 'offhand' argument comes very early in *Theory*, and anticipates a more complex argument to come in support of its conclusion, which may yet prove able to support it. But Rawls had already provided a brief description of his original position and veil of ignorance, and in the light of that description, it seems quite likely, 'offhand', that the parties would agree to a principle, such as the principle of average utility, which may require lesser life prospects for the sake of a greater sum of advantages enjoyed by 'others'. This would be more likely to be the case given that, since they would be agreeing to the principle of average utility, those 'others' may turn out to be themselves.

Reconstruction of Rawls's argument of Theory

1.9.44 The next question to ask is whether Rawls's 'offhand' argument of *Theory* would be any more promising if it were reconstructed along similar lines to the 'more promising' reconstruction of his argument of 'Distributive Justice'?

1.9.45 And the simple answer is: 'No'. A similar reconstruction of Rawls's argument is, in fact, far less promising from the point of view of Justice as Reciprocity than the reconstructed argument of 'Distributive Justice'. What made the reconstructed argument of 'Distributive Justice' promising, was, I claimed (in §1.9.20) the fact that it would be rational for someone who knew he would be a 'loser' in society to veto any principle that made him worse off than retaining 'the liberty of the original position'. But no analogous

line of argument would make it clearly rational for a loser, even one who knew he would be such, to veto a principle that may require him to have lesser life prospects imply for the sake of the greater advantages of others, as the following paragraphs (§§1.9.46 - 1.9.50) explain.

1.9.46 In fact, there are two analogous lines of argument to consider, caused by the substitution of the italicized part of [B] for its equivalent in **Passage 1k *(DJ 1967),*** [2]. The effect of this substitution is to cast confusion over the questions of exactly who are those entities who may be required by the principle of utility to have 'lesser life prospects' for the sake of the greater advantages of others, and lesser prospects than what? Is it the parties in the original position who may be required to have lesser life prospects than if they retained the liberty of the state of nature? Or is it the losers in the well-ordered society who may be required to have greater life prospects for the sake of advantages to the winners in the well-ordered society?[1]

1.9.47 Let us take the first of these possibilities first. Suppose we read the claim as asserting that the parties in *the original position*, if they chose the principle of utility, would require the losers in the well-ordered society to have less than they would in a state of general egoism for the sake of the winners. This is not only an unnatural reading of the claim, but is contradicted by **Passage 1s (*T of J Rev*)[2],** which, as just remarked (§1.9.36), asserts that the principle of utility would mean that *all*, including the worst off in society, would enjoy greater life prospects than in a state of general egoism. An insinuation that some would veto the principle of utility because it required them to have lesser life prospects than in a state of general egoism would simply be false.

1.9.48 It would be more natural to interpret the claim as being confined to the citizens of a well-ordered society, and this is how it has been sometimes interpreted.[3] This raises its own difficulties. As Thomas Nagel and Alan Gibbard have pointed out, a similar charge could

[1] Robert Nozick (1974 p. 196) makes a similar point regarding the ambiguity of another passage of *Theory orig* over whether it is referring to the parties in the original position or the citizens of a well-ordered society.
[2] p. 67.
[3] For example, see Scanlon 'Contractualism and Utilitarianism' (Scanlon 1982 p.123 fn.18).

be pressed against the two principles of justice in comparison with the principle of utility.[1] The two principles of justice would require lesser life prospects for those who would do better under the principle of utility for the sake of greater advantages to those who would do better under the two principles of justice. For this reason the objection has generally been interpreted as an objection to sacrifices of the worse off for the better off, and in the revised edition of *A Theory of Justice*, Rawls appears to have clarified that the objection should be so interpreted.[2]

1.9.49 So it seems reasonable, henceforth, to interpret **Passage 1j *(T of J Rev)***[3] as objecting to sacrifices of the less advantaged for the sake of the more advantaged. On this interpretation, the reconstructed argument claims that losers would veto the principle of utility if it may require them to have lesser life prospects for the sake of the greater advantages of others once they 'arrive' in the well-ordered society. But such a veto would only make sense if they could expect to hold out for something better. It would be foolish of them to exercise the veto if that would scupper the deal, and leave them where they started. It would be more rational for them to choose to endure making sacrifices for the winners in a society that still offered them better prospects than in a state of general egoism.

1.9.50 The key claim of Sentence [B] **Passage 1j *(T of J Rev)***, then, that it seems unlikely that parties would agree to the principle of utility, *even on the reconstructed argument where those who would turn out to be losers knew as much*, looks unsustainable whichever way **Passage 1j *(T of J Rev)*** is read.

1.9.51 The remaining claims of **Passage 1j *(T of J Rev)*** are also highly questionable. Rather than go through all of them, I shall just compare [D]'s claim that 'a rational man would not accept a basic structure merely because it maximized the algebraic sum of advantages irrespective of its permanent effects on his own basic rights and interests' and

[1] See Nagel 'Rawls and Justice' (Nagel 1973 p.13) and Gibbard *Reconciling our Aims* (Gibbard 2008 pp.40 -41)
[2] See Chapter 3, Section 2 of this thesis.
[3] p. 68

compare it with its correlative from **Passage 1k *(DJ).*** That correlative was sentence [4]'s 'A rational man would not accept an institution merely because it maximized the sum of advantages irrespective of its effect on his own interests.' So it is very similar. It made sense, however, in the context of 'Distributive Justice' to assert that a rational man would not accept an institution ordered by the principle of utility when his interests may be better served by the liberty of the original position, merely to maximize the sum of advantages. It makes much less sense, in the context of *Theory,* to assert that 'a rational man would not accept a basic structure merely because it maximized the algebraic sum of advantages irrespective of its permanent effects on his own basic rights and interests.' If the fallback position in the event of non-agreement was a state of general egoism, a man who chose the principle of utility in preference to a state of general egoism would not be accepting a basic structure merely because it maximized the algebraic sum of advantages irrespective of its permanent effects on his own basic rights and interests. He would be better described as accepting a basic structure that maximized the algebraic sum of advantages because doing so would likely secure him some basic rights and enable him to effectively pursue his interests.

1.9.52 There may well be some other ways of interpreting the charges against utilitarianism contained within these passages that I haven't considered. But I doubt very much that they would prove more coherent. For the truth is, I think, that the **Passage 1j *(TJ Rev)*** from *Theory* can be explained as an attempt to plaster over the cracks in the argument of **Passage 1k *(DJ 1967)*** of 'Distributive Justice'. But the need to fix that argument was forced by a substantive change in Rawls's position. An underlying assumption of the argument of **Passage 1k *(DJ 1967)*** was that the principle of utility would fail to compensate all cooperating members of society for their loss of the liberty of the original position. But by the time of *Theory*, Rawls was committed to the position that the utilitarian conceptions of justice *would* improve the position of all cooperating parties in comparison to the equivalent liberty of the original position, which was a state of general egoism. New cracks in the plaster duly appeared, and the only filler available to Rawls was to claim that the principle of utility might, in contrast to the two principles of justice, require the less advantaged to make sacrifices for a greater sum of benefits to the more advantaged. From the perspective of Justice as Reciprocity this does not, at initial

inspection, look like such a damaging charge as that the principle of utility would fail to provide all with compensation for their loss of natural liberty.

Concluding Remarks

1.10.1 The comparison of **Passage 1j *(T of J Rev)*** from *Theory* with its historical antecedent **Passage 1k *(DJ 1967)*** from 'Distributive Justice' has, I hope, exposed two major, and related problems for Rawls's theory of Justice as Reciprocity. The first problem is that Justice as Reciprocity might prove to be reconcilable with utilitarianism after all. The second is that, just because Justice as Reciprocity might be reconcilable with utilitarianism, it might fail to support the three tenets of deontological liberalism.[1]

1.10.2 I suggested above (§§1.6.1 - 1.6.7) that the great appeal of Rawls's theory of justice was that it offered the promise of underwriting 'common sense' intuitions about justice, that correspond to the three tenets of deontological liberalism, with a plausible account of their moral foundation in the conception of Justice as Reciprocity. Rawls envisaged this could be done through the route shown on the left hand side of Fig. 1(i). Justice as Reciprocity would entail Justice as Fairness, which would in turn entail the two principles of justice, which would in turn uphold the three tenets of deontological liberalism.

1.10.3 But when Rawls first came up with his 'idea' for his theory of justice [2] he was committed to assumptions that would make that route work.[3] By the time of *Theory* his assumptions had changed, and the route could quite conceivably lead to acceptance of the principle of utility, in either its average or classical version. This contention can be confirmed by observing that a modified Justice as Fairness could support the principle of

[1] I explained utilitarianism's inability to support the three tenets of deontological liberalism above (§§1.2.1 – 1.2.3)
[2] Wolff 1977 p.16 describes Rawls's as 'one of the loveliest ideas in the history of social and political theory', though he interprets Rawls's idea somewhat differently to myself.
[3] Though, as I shall argue in Chapter 2, the stage of Justice as Fairness was actually an unnecessary diversion when a direct route from Justice as Reciprocity to the two principles was viable.

utility, with a tinkered veil of ignorance which gave those who would be likely to be the worst off group in society the opportunity to reject the principle of utility in favour of 'the liberty of the original position'.

1.10.4 For imagine you were faced with a choice today between accepting the rule of a potential Leviathan[1], who happened to be a classical utilitarian, or the 'liberty of the original position' which I suggested above (§§1.9.26 - 1.9.28) could be coherently interpreted on roughly Hobbesian lines as the right to do anything in pursuit of whatever aims one happens to have. You know that you will be in the worst off group in society. You also know that a classical utilitarian Leviathan is quite prepared to countenance doing extremely nasty things to members of the worst off group in society for the sake of greater advantages to other groups in society. For example, the classical utilitarian Leviathan would be prepared to institute a system of slavery so abject that the slaves would consider themselves better off dead, if to do so would promote the greatest happiness. Such an abject system of slavery would make your life more unpleasant than it would be if you retained the 'liberty of the original position'. On top of that, she would insist that all slaves, including yourself, had an obligation to obey all the duties of their station, even though doing so would afford them no prospect of a life worth living. And you know that you would be one of those slaves. Finally, she would not be tolerant of alternative ideologies or lifestyles. She would insist that classical utilitarianism was the one true way and that all the citizens in in her society should follow it.

1.10.5 However, you also know that, for reasons such as diminishing marginal utility, outlined in §1.5.2 above, the Leviathan would be far more likely to arrange institutions in a similar way to a Leviathan who accepted Rawls's two principles of justice. The Leviathan has, after all, read John Stuart Mill's *On Liberty* and believes that the rights to safeguard the individual from the state should be taken very seriously indeed, even if they are not inviolable. Having read *On Liberty*, she also believes that it is more practically sensible to leave people to find their own path to the one true way of Classical Utilitarianism than to force it on them.

[1]The 'Leviathan' of Hobbes's title refers to a ruler or government.

1.10.6 You happen not only to be one of those who are going to be a member of the worst –off group in society, but a member of the quasi-religious sect of 'Prioritarians'. Prioritarianism holds that everyone should aim at making the worst-off group in society as well-off as possible. A Leviathan who was committed to Rawls's two principles of justice would, then, be more to your personal advantage and to your ideological taste. He would be more to your personal advantage because you would be better off under his jurisdiction and he would be more to your ideological taste because he would be committed to making the worst-off group in society as well off as possible. Unfortunately, a Leviathan who is committed to Rawls's principles of justice is not available.

1.10.7 In this modified original position it would, I maintain, be rational for everyone, including those who knew they would be in the worst off group in society (and including Prioritarians) to agree to a contract pledging obedience to the classical utilitarian Leviathan. Then, if we adopted this modified original position as 'the appropriate initial status quo' of Rawls's definition of Justice as Fairness given in **Passage 1f (*T of J Rev)*[1]**, Justice as Reciprocity would have proceeded, via Justice as Fairness, to have instituted classical utilitarianism.[2]

1.10.8 In this case, Justice as Reciprocity would not only have failed to repudiate utilitarianism, but it would also have failed to deliver on the promise to underwrite the three tenets of deontological liberalism (§1.1.1 & §§1.6.1 – 1.6.7). It would have failed to underwrite them firstly, because no-one would have any right not to act so as to maximize utility, secondly; because everyone would have the right to do anything to do anything to anyone else for the sake of the maximization of utility and thirdly; because no one would have the right to choose any alternative conception of the good than to be a classical utilitarian.

1.10.9 This, I think, is a fair summary of the problem faced by Rawls which, I shall argue in the forthcoming chapters, he never surmounted.

---

[1] p. 45

[2] This route is represented by the wiggly line in Fig. 1(i)

Chapter 2. The First Model of Justice as Fairness

2.0.1 The concluding remarks of Chapter 1 remarked that when Rawls first came up with his idea of Justice as Reciprocity he was committed to assumptions that would make the route from Justice as Reciprocity to his two principles of justice work (§1.10.3). This would have had the added bonus of fulfilling the promise of Justice as Reciprocity to support the three tenets of deontological liberalism. The point of examining Rawls's first model of Justice as Fairness in this chapter in close detail is to demonstrate the advantages of Rawls's first model in that regard. This will set the scene for the rest of this thesis which argues that once Rawls abandoned the assumptions of the first model he was unable to find an alternative route that worked.

2.0.2 Rawls's first model did not only have the advantage of establishing the route from Justice as Reciprocity, it would also have established the two principles of justice as the *only* conception of justice capable of meeting the requirements of Justice as Reciprocity, set out in Chapter 1 (§§1.4.1 - 1.4.2). This is because they would have been the *only* principles capable of meeting **the mutual advantage condition**. And by meeting the mutual advantage condition they would meet **the fairness condition.** It is another important aim of this chapter to show that, on Rawls's first model, meeting the mutual advantage condition was a necessary and sufficient condition of meeting the fairness condition.

2.0.3 The repercussion of the success of Rawls's argument of his first model would have been to erect insurmountable blockages on both the direct route from Justice as Reciprocity to the utilitarian conceptions of justice, and the route proceeding via Justice as Fairness. The utilitarian conceptions of justice would have been unable to meet the mutual advantage condition, and by failing to meet the mutual advantage condition they would be equally incapable of meeting the fairness condition.

2.0.4 Rawls's first model would also have had the advantage of being able to sustain all three of the tenets of deontological liberalism

2.0.5 But, as I show in this chapter, the repercussions of the failure of Rawls's first model

removed at least part of the blockages on the routes from Justice as Reciprocity to the utilitarian conceptions of justice. At the end of the chapter I will show that the utilitarian conceptions would have been able to meet the mutual advantage condition, given Rawls's abandonment of many of the assumptions that underpinned the first model, and the revised assumptions of *Theory*. The major revision which would enable the utilitarian conceptions of justice to meet the mutual advantage condition, was Rawls's shift to defining mutual advantage in comparison to the benchmark of a Hobbesian state of nature.[1] The end of this chapter will leave the utilitarian conceptions able to meet the constraint requirement and mutual advantage condition of Justice as Reciprocity with a question still hanging over their ability to meet the fairness condition. It will also call into question the ability of Justice as Reciprocity to uphold the three tenets of deontological liberalism.

2.0.6 Rawls first put forward his two principles of justice in his 1957 paper 'Justice as Fairness', which was to be presented at a Symposium later the same year, and was published in the Journal of Philosophy. A longer essay of the same name was published the following year, and it is the fuller version which has generally been treated as Rawls's first statement of his theory of Justice as Fairness.[2] I shall refer to both essays, which do not contain any important theoretical differences, but will pay rather more attention to the earlier essay than most other commentators have seen fit to do. This is primarily because it was in the first version of 'Justice as Fairness' (henceforth to be referred to as 'Justice as Fairness (1)') that Rawls explicitly referred to his theory as one of 'Justice as Reciprocity' which he did not do in the second version (henceforth to be referred to as 'Justice as Fairness (2)')

2.0.7 Robert Paul Wolff remarks of what he variously describes as 'the first form of the model' or 'the first model' as presented in those two essays: 'In its first form, Rawls's model is simple, clear, elegant, and - as we shall see - subject to devastating objections. Despite its shortcomings, however, the first form of the model is, I will argue, the real

[1] As discussed at length in my Chapter 1 (§§1.9.23 – 1.9.40)
[2] For example, Robert Paul Wolff focuses exclusively on the 1958 essay in his discussion of 'the first form of Rawls's model' and Samuel Freeman made the editorial decision to omit the initial version in Rawls's *Collected Papers* (1999)

foundation on which all the rest of Rawls's theory is constructed.'[1]

2.0.8 Wolff's remark concisely summarizes my view of Rawls's first model. I also believe that the interpretation Wolff gives of Rawls's two principles as they were in Rawls's first model is correct, as is his account of the first form of Justice as Fairness. So here I shall not add much to Wolff's exegesis of Rawls's theory as it stood then, apart from drawing attention to aspects particularly relevant to the conception of Justice as Reciprocity and Rawls's critique of utilitarianism.

In **Section 1**, I describe Rawls's principles of justice as they should be understood in the first model, and draw particular attention to the first model's argument against 'aggregation' – weighing advantages to some against disadvantages to others – that utilitarianism is unavoidably committed to. Many philosophers have objected to utilitarian aggregation, as it is the aggregative feature of utilitarianism that makes it unable to accommodate the first two tenets of deontological liberalism. I go into this argument against the purported injustice of utilitarianism in more detail in the section on 'the separateness of persons' objection to utilitarianism in Chapter 4, but that later argument will only make sense against the backdrop of the argument against aggregation that Rawls would have been able to put had his first model proved viable. The argument of Sections 1 to 4 will proceed on the assumption that Rawls was committed to what I call '**the simple assumption of the first model**'.

Section 2 gives an account of Rawls's theory of Justice as Fairness, as it stood then. This section will, as the first section does, largely follow Wolff's account from *Understanding Rawls*, though I shall put forward an important argument of my own to support the idea that in the event of a clash, Justice as Fairness (that is, the stipulation that those principles which would be chosen in the appropriately defined original position are those that should be considered 'just') should give way to the requirement that all people who have a duty to cooperate in the well-ordered society should be advantaged with respect to a position of equal liberty. This argument will help pave the way for a further argument against Justice

[1]Wolff 1977 p.25

as Fairness that I put in Chapter 3.

In **Section 3**, I shall argue that Hart's 'principle of fair play' – or the 'duty of fair play' as Rawls referred to it then - played a far more important role in Rawls's first form of the model than it subsequently did in *Theory*. Indeed, it stipulated conditions that only the two principles of justice could hope to meet. By meeting them, the two principles of justice would be the only principles that could possibly be suitable for the conception of Justice as Reciprocity. Neatly enough, the principle of fair play maps exactly on to Gibbard's defining features of reciprocity that I set out in Chapter 1 (§§1.4.1 -1.4.2). Since the duty of fair play was, on Rawls's first model, only compatible with the two principles of justice it had the effect of excluding the utilitarian conceptions of justice from consideration. Importantly, from the point of view of this thesis, neither the wiggly line nor the 'bypass 2' route shown in Fig. 1(i) would have been remotely viable.

In **Section 4**, I examine Rawls's first use of the term 'reciprocity' in 'Justice as Fairness (1)', and his conception of Justice as Reciprocity. I also examine Rawls's early conception of Justice as Benevolence and its relationship to utilitarianism, and show how the idea that the two would advocate the same principles of justice appeared to be precarious.

Section 5 considers the possibility that Rawls was not committed to the simple assumption of the first model that I attributed to him in Section 1. It examines an argument of Brian Barry's that purports to show that Rawls's difference principle is the most egalitarian, or impartial, compromise on straight equality in the face of the problem posed by there being more than one distribution that is Pareto preferred to equality. I suggest that Barry's argument fails, but also that the success or failure of an argument establishing the difference principle as the egalitarian alternative to equality would have made little difference to the first model's argument that the aggregation that utilitarianism is unavoidably committed to was unjust.

In **Section 6**'s concluding remarks I consider the potential implications of Rawls's revision of the assumptions of his first model for his enduring contention that the two principles of justice are grounded in the conception of reciprocity whilst utilitarianism is grounded in

benevolence, and never the twain shall meet. I also consider the effect of the failure of Rawls's first model on the 'promise', described in Chapter 1 to support the three tenets of deontological liberalism with the conception of Justice as Reciprocity.

1 The two principles of justice

2.1.1 There are two radical differences between the first model and the model of *Theory* that are especially important from the point of view of my argument. The first difference is as follows. In *Theory*, the two principles of justice are concerned with different things: the first is concerned with the basic liberties and the second is concerned with economic distribution. In the first model, by contrast, the two principles were continuous with each other and both concerned with *both* liberty *and* economic distribution. The second difference is that 'liberty' was measured in a very different way to how it would later be measured in Theory; it was to be measured in terms of the economic advantages that different assignments of rights would bestow on the holder of those rights. Rawls did not explain these points very clearly in either 'Justice as Fairness' (1) or 'Justice as Fairness' (2), so some analysis of the texts of those essays is needed here to show that the two principles in the first model should indeed be interpreted in the way I have just suggested.

2.1.2 In 'Justice as Fairness' (2), Rawls wrote

Passage 2a (*J as F 2)*

> The conception of justice which I want to develop may be stated in the form of two principles as follows: first, each person participating in a practice, or affected by it, has an equal right to the most extensive liberty compatible with a like liberty for all; and second, inequalities are arbitrary unless it is reasonable to expect that they will work out for everyone's advantage, and provided the positions to which they attach, or from which they may be gained, are open to all. These principles express justice as a complex of three ideas: liberty, equality, and reward for services contributing to the common good. [1]

[1] Rawls 1958 p.165-166

2.1.3 Rawls did not explain his notion of practice in any detail, simply remarking in a footnote that

Passage 2b (*J as F 2)*

> I use the word "practice" throughout as a sort of technical term meaning any form of activity specified in a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure. As examples one may think of games and rituals, trials and parliaments, markets and systems of property.[1]

2.1.4 The two passages just cited describe how the first principle requires an equal liberty (as the first principle of justice in *Theory* would still do) and the second allowed inequalities that were to everyone's advantage provided there was equal opportunity to achieve the advantageous positions (as the second principle of justice in *Theory* would still do). But the description of the principles given so far does not show that they are both supposed to be concerned with liberty. Nor do the passages indicate how 'liberty' should be understood or measured. The passage which provides evidence that Rawls then regarded both principles to refer to liberty, understood as an assignment of rights to people in different positions, and for any inequality in liberty to be measured in gains in material or other quantities, is

Passage 2c (*J as F 2)*

> The second principle defines what sort of inequalities are permissible; it specifies how the presumption laid down by the first principle may be put aside. Now by inequalities it is best to understand not *any* differences between offices and positions, but differences in the benefits and burdens attached to them either directly or indirectly, such as prestige and wealth, or liability to taxation and compulsory services. Players in a game do not protest against there being different positions, such as batter, pitcher, catcher and the like, nor to there being various privileges and powers as specified by the rules: nor do the citizens of a country object to there being the different offices of government such as president, senator, governor, judge, and so on, each with their special rights and duties…they may object to the distribution of power and wealth which results from the various ways in which men avail

[1] Rawls 1958 fn.2 p. 164

> themselves of the opportunities allowed by it (e.g., the concentration of wealth which may develop in a free price system allowing large entrepreneurial or speculative gains).[1]

2.1.5 So the idea is that just as players in different positions in a game of baseball have to abide by different rules governing how they play the game (e.g. a batter is not allowed to catch a ball, an outfielder is) so citizens participating in the practices of society may have different rights assigned to them determining what they are permitted or not permitted to do. The second principle of justice allows different assignments of rights to the occupants of different roles in the practices of society providing everyone benefits with respect to what they would have under the same assignment of rights. The inequalities allowed by the second principle of justice, then, refer to different assignments of rights – or 'liberty' – and the value of the different assignment of rights is to be measured in terms of material or psychological gains. I cannot improve on Wolff's summary of the 'central idea' of Rawls's principles, which he puts thus:

Passage 2d (*Wolff 1977)*

> The central idea behind Rawls's principles seems clear enough: the output or earnings of a practice is to be distributed equally, unless some pattern of unequal distribution can…be made to work for everyone's benefit, and provided that everyone has a shot at the better-paid roles.[2]

2.1.6 This interpretation of Rawls's principles as both applying to liberty, understood in terms of assignments of rights, with the advantages of various assignments of rights being measured essentially in economic terms calls for some justification. Once again I follow Wolff. Wolff acknowledges the fact that Rawls's formulation of his principles as concerned with 'liberty' is 'puzzling' when they appear to be more concerned with economic distribution, but supports his reading of Rawls's principles with two considerations. The first is textual. In reply to an anticipated objection that Rawls should instead be read as applying his first principle to the issue of political liberty, and the second to economic distribution, as he was to go on to apply them in *Theory*, Wolff points out that

[1] Rawls 1958 p.167
[2] Wolff 1977 p.38

Passage 2e (*Wolff 1977)*

> Rawls does not say that the second principle specifies the conditions under which political liberty may be set aside for economic advantage. He does not, in other words, present the second principle as grounds for overriding the first. Rather, he presents the second principle as stating the grounds on which the presumption (of equal distribution) can be set aside. What is at stake, quite clearly, is the question when unequal distribution of payoffs may justly be substituted for equal distribution of payoffs, *not* the quite different question when a certain pattern of payoffs of one sort of good (wealth, etc) may be invoked as justification for deviating from an equal distribution of a different sort of good (namely, liberty).[1]

2.1.7 Wolff's reasoning here appears to me to be sound. His point is that Rawls did not write something along the lines of, "Everybody should have the greatest equal liberty compatible with a similar equal liberty for all, unless equal liberty interferes with economic advantage, in which case liberty can be sacrificed for the sake of the pursuit of economic advantage." Instead of which he wrote that the presumption of equality could be waived in favour of inequality, presumably applied to the same thing. Wolff's reading can be supported by the observation that **Passage 2c (*J as F 2)***[2] envisages the inequalities that may be permitted by the second principle being attached to 'different offices' with 'special rights and duties'. So the idea seems to be that the second principle of justice will allow some people to have advantages in comparison to others by virtue of the 'special rights' they have, and that the possession of 'special rights' is equivalent to having 'extra liberty'.

2.1.8 The second consideration offered by Wolff in favour of his interpreting the second principle of justice as being applied to liberty is that such an interpretation

Passage 2f (*Wolff 1977)*

> makes Rawls's "theorem" seem at least initially plausible. The proof of the theorem will simply involve an invocation of the conception of Pareto

[1] Wolff 1977 p. 39
[2] p. 82

> optimality, with the understanding that the quasi-ordering of alternative distributions is to be made with respect to the original, or baseline situation of equal distribution. Such a line of reasoning in support of the two principles will make sense only if the first principle is construed as a prima facie rule of equal distribution and the second principle is construed as an excuse for deviations from distributive equality.[1]

2.1.9 The passage above provides a useful bridge from the question of how Rawls's principles of justice in the first model should be interpreted to the question of what the implications of the first model would be for the permissibility of any attempts to aggregate advantages, in the way that the classical principle of utility would be apt to do. I believe that the Rawls of the first model as represented in 'Justice as Fairness' (1) and (2) should be read as committed to **the simple assumption of the first model**, which would exclude aggregation, the 'weighing of advantages to some against disadvantages to others', *however* that ambiguous phrase might be interpreted, and the first ground I shall offer for that belief is similar to the consideration offered by Wolff in **Passage 2f (*Wolff 1977*).**[2] As Wolff pointed out, Rawls appears to 'simply' invoke the conception of Pareto-optimality in support of his principles. He does not, in other words, consider the important question of which distribution the two principles would select if there were two or more Pareto-optimal distributions none of which were Pareto-preferred to any other. This suggests to me that he assumed that there would be *at most, only one way* in which a practice satisfying Rawls's first principle of justice could be altered and meet the requirements of his second principle of justice. To put the point another way: there could be only one Pareto-optimal distribution that is Pareto-preferred to equality.

2.1.10 The points of the previous paragraph, and the technical terms involved (i.e. 'Pareto-preferred', 'Pareto-optimal', 'aggregation' and 'the simple assumption of the first model'), will make more sense in the light of the discussions I am about to give of first, a putative counter-example to the difference principle and secondly, an important passage from 'Justice as Fairness' (2).

2.1.11 The counter-example is based on one given by John Broome in an appendix to

[1] Wolff 1977 p.40
[2] I provide more grounds for this belief below.

Derek Parfit's *Reasons and Persons*, adapted to be better suited to Rawls's first model.[1] It will be subject to several variations over the next few chapters to reflect the changing assumptions of Rawls's models and different problems to be illustrated. Let us suppose, then, that there exists a country in 2016, 'Freedonia' with a population divided equally between agricultural and non-agricultural workers. Suppose that without price supports for agricultural product the agricultural workers would be the worst off group in Freedonia – price supports would make them better off. However, price supports would have the effect of worsening the economic position of everyone else, but not enough so as to make them as badly off as the agricultural workers would be without price supports. The situation is illustrated in the figure below. The numbers in the table can be taken to represent 'net benefits' so as to be consistent with the 'benefits' and 'burdens' referred to in **Passage 2c *(J as F 2)***[2]

Fig 2 (i)[3]

Distribution	Non-agricultural workers	Agricultural workers
1	115	140
2	120	110
3	100	100

2.1.12 The difference principle, as justified by the use of the maximin principle in *Theory,* would select Distribution 1 as that is the distribution where the 'worst off group' is as well off as possible. But what would the first model's second principle of justice recommend? The second principle of justice, as described in 'Justice as Fairness' (1&2) and as outlined in the last few sections, would seem to be compatible with both Distributions 1 & 2 as they are both unequal distributions that are advantageous to everyone in comparison to the equal

[1] Joshua Cohen uses the example of price support for agricultural workers in his essay 'Democratic Equality' (Cohen 1989 p. 739). This suits Rawls's first model as Rawls specified 'markets' as an example of a practice in **Passage 2b (*J as F 2)*** p. 82**.** Broome's example described was aimed at Rawls's third model of *Theory,* which applied the principles of justice to the basic structure of society. So Broome's example imagined different constitutions India and Britain might have in 1800 (Parfit pp. 491-492).

[2] p. 82.

[3] All the examples in this chapter assume that Freedonia has a fixed population and the recommendations of the classical principle of utility and the principle of average utility would coincide. So it is convenient just to talk of 'utilitarianism' and 'the principle of utility'.

Distribution 1.

2.1.13 The question as to which of the two unequal distributions the second principle of justice would select has arisen because my example has supposed that there are two distributions that are ***Pareto-preferred*** to equality. A distribution, *x*, is weakly Pareto-preferred to a distribution, *y*, if at least one person prefers *x* to *y* and no-one prefers *y* to *x*. It is strongly Pareto-preferred to equality if everyone prefers *x* to *y*. As Wolff explains, and as shown in **Passage 2g (*J as F 2***) below[1], the principle of Pareto preference that Rawls invokes is that of strong Pareto preference rather than weak Pareto preference.[2] Both Distributions 1&2 are ***Pareto-optimal.*** A distribution is Pareto-optimal if it is impossible to make anyone better off without making someone else worse off. If Distribution 1 were chosen rather than Distribution 2, the agricultural workers would be better off but everyone else would be worse off. The converse would be true if Distribution 2 were chosen rather than Distribution 1.

2.1.14 Now, if Rawls were committed to ***the simple assumption of the first model***, this question would not arise. This assumption would have the effect of limiting the feasible set to either Distributions 1&3 or Distributions 2&3. Suppose it were limited to 1&3. Then we would have

Fig 2 (ii)

Distribution	Non – agricultural workers	Agricultural workers
1	115	140
3	100	100

2.1.15 The second principle of justice would unequivocally be in favour of Distribution 1, as that distribution would be preferred by all to 3, and would be the only distribution that was preferred by all to 3.

[1] p. 88
[2] Wolff argues convincingly that Rawls first form of the model should have only committed him to weak Pareto-preference. See Wolff (1977) pp 40-41.

2.1.16 So my first ground for ascribing the simple assumption of the first model to Rawls is that it would avoid any dilemma as to how to *apply* the second principle of justice in the case of there being more than one Pareto-optimal unequal distribution. If Rawls was aware of the problem, he certainly never addressed it in either 'Justice as Fairness' (1) or (2).

2.1.17 My second ground for ascribing the simple assumption of the first model to Rawls is that it would avoid a dilemma as to how to *interpret* the following passage which contains Rawls's first objection to utilitarian **aggregation**.

Passage 2g (*J as F 2*)

> [1] It should be noted that the second principle holds that an inequality is allowed only if there is reason to believe that the practice with the inequality, or resulting in it, will work for the advantage of *every* party engaging in it. [2] Here it is important to stress that *every* party must gain from the inequality. [3] Since the principle applies to practices, it implies that the representative man in every office or position defined by a practice, when he views it as a going concern, must find it reasonable to prefer his condition and prospects with the inequality to what they would be under the practice without it. [4] The principle excludes, therefore, the justification of inequalities on the grounds that the disadvantages of those in one position are outweighed by the greater advantages of those in another position. [5] This rather simple restriction is the main modification I wish to make in the utilitarian principle as usually understood. [6] When coupled with the notion of a practice, it is a restriction of consequence, and one which some utilitarians, e.g., Hume and Mill, have used in their discussions of justice without realizing apparently its significance, or at least without calling attention to it. Why it is a significant modification of principle, changing one's conception of justice entirely, the whole of my argument will show.[1]
> [My line numbering, Rawls's italics]

2.1.17 **Passage 2g (*J as F 2*)** starts out, in Sentence [1], by interpreting the second principle of justice as requiring that practices with inequalities are allowed only if they work for the advantage of *everyone* engaged in them. Sentence [2] just reiterates the same

[1] Rawls 1958 pp 67- 68

point. Because Rawls has asserted that it is only if a practice meets this requirement that the equality requirement of the first principle can be waived, he presumably means practices to at least be to everyone's advantage in comparison to the same practice under equality. I shall call that **requirement (a)**. 'Every' is stressed in each of the first three sentences, providing confirmation of Wolff's claim that Rawls meant to invoke strong, rather than weak, Pareto preference over equality as a requirement of his second principle of justice. Requirement (a) is, I think, unambiguous.

2.1.18 Sentences [3] and [4] are ambiguous, and it is their ambiguity that leads to the differing possible interpretations of this passage. When, in Sentence [3], Rawls stipulated that every representative man 'must find it reasonable to prefer his condition and prospects with the inequality to what they would be without it' – I shall call this **requirement (b) -** did he mean i) that they must prefer their condition and prospects to their condition and prospects under *any other possible configuration* of the practice that met requirement (a) including all unequal ones that are Pareto-preferred to equality, or did he just mean ii) that they must prefer their condition and prospects with the inequality to their condition and prospects under equality? The ambiguity of Sentence [3] carries over into Sentence [4]. Sentence [4] excludes any inequalities that might potentially be justified on the grounds that the greater advantages of some outweigh the lesser disadvantages of others. This is the first variant of Rawls's extremely influential claim that distributive justice excludes aggregation. I shall refer to 'the exclusion of aggregation' expressed by Sentence [4] **requirement (c)**.[1] But did Rawls intend 'disadvantages' and 'advantages' to be i) measured by reference to *any other possible configuration* of the practice that would be Pareto-preferred to equality, or ii) just to the practice under equality?[2]

2.1.19 This problem of interpretation can be clarified with reference to **Fig 2 (i)** and **Fig 2 (ii).** First, let us suppose that Rawls did not make the simple assumption of the first model, so the feasible set might be that of **Fig 2 (i).**

2.1.20 A sensible way to resolve the problem of how to decide between distributions 1 & 2

[1] I shall refer back to the 'exclusion of aggregation' in Chapter 4.
[2] The two interpretations shall henceforth be referred to as interpretations (i) and(ii).

in Fig 2 (i) might appear to be to see where the greater sum of advantages lay and pick Distribution 1, if there were a sufficient number of agricultural workers for that choice to maximize the sum of advantages or Distribution 2, if that choice would maximize the sum of advantages. To aggregate, in other words. But would choosing on such grounds be a violation of requirements (b) and (c)? The answer is that it would be, on interpretation (i) of those requirements, but wouldn't be on interpretation (ii) of those requirements.

2.1.21 This dilemma of interpretation would not arise, however, if Rawls was committed to the simple assumption of the first model. In **Fig 2 (ii)** there is only one unequal distribution that is Pareto-preferred to equality, Distribution 3, and that distribution is would satisfy requirements (b) and (c) on both interpretations (i) and (ii) of those requirements. That provides my second ground for supposing that Rawls was committed to the simple assumption of the first model. It would contain an unequivocal objection to any aggregation, however advantage and disadvantage might be understood. And just as Rawls did not consider the question of how to resolve the problem of how to apply the two principles of justice in case there were more than one unequal distribution that was Pareto-preferred to equality, so he did not consider the question of how to interpret requirements (b) and (c) in 'Justice as Fairness' (2).

2.1.22 A third ground for ascribing the simple assumption of the first model to the Rawls is that he later *did* address the problem posed by the possibility there being more than one unequal distribution that is Pareto-preferred to equality in his 1967 essay 'Distributive Justice', as if it had occurred to him for the first time. It is not mentioned in the three essays in between. Wolff takes 'Distributive Justice' to be the defining essay of the second form of the model.[1] The innovations introduced in that essay, in particular the concepts of chain connection and close-knitness, can, I think, be understood as motivated largely by the desire to surmount the problem arising from the possibility of there being more than one unequal distribution that is Pareto-preferred to equality.[2]

2.1.23 Three more reasons for supposing that Rawls was committed to the simple

[1] Wolff 1977 p. 5
[2] This point will receive further explanation in Chapter 3

assumption of the first model are first, the advantage it would bestow to Rawls's ambition to uphold the two principles of justice as the principles of justice that were particularly suited to the conception of Justice as Reciprocity, secondly, the advantage it afforded to Rawls's ambition to refute utilitarianism and thirdly, the neatness of fit it would provide to the different elements of his theory at the time. The rest of this section will consider the first two of these advantages. The third should reveal itself over the analysis of the different elements of Rawls's first model in Sections 2 to 4. The aim of demonstrating this neatness of fit justifies my proceeding on the assumption that Rawls was committed to the simple assumption of the first model before going on to question this in Section 5.

2.1.24 The advantage the simple assumption of the first model would bestow to Rawls's ambition to uphold the two principles of justice as the principles of justice that were particularly suited to the conception of Justice as Reciprocity should now be easy to understand. In Chapter 1, I suggested that a necessary condition for principles of distributive justice to be suited to that conception was that they meet the **mutual advantage condition,** whatever that may turn out to involve. The benchmark by which to measure mutual advantage could be taken to be 'equal liberty'. The two principles of justice would then emerge as *uniquely* suited to the conception of Justice as Reciprocity. The first principle of justice would ensure that all cooperating members of society benefited with respect to what they would have in a distribution that gave them any less than the greatest possible equal liberty.[1] The second principle of justice would then ensure that any unequal distribution that was to the advantage of all cooperating parties would be selected. Given the simple assumption of the first model there would be no need for any other criteria to decide which mutually advantageous distribution should be selected. So the two principles of justice would appear to be the principles that best satisfied **the mutual advantage condition**. Any distribution that didn't satisfy them would be to *everyone'*s disadvantage in comparison with the unique distribution that did satisfy them.

2.1.25 The simple assumption of the first model would also have had the effect of making

[1] Recall that 'equal liberty' in the first model is to be understood in terms of assignment of rights that bestow equal advantages on people. Presumably Rawls envisaged there being a potential assignment of rights that bestowed equal advantages on all, but less than the maximum advantage compatible with equality.

a very powerful case that any aggregation of the kind that utilitarianism, in any of its forms, is committed to, would be unjust according to the lights of Justice as Reciprocity. To appreciate this, we first need to note that the simple assumption of the first model only asserts that all unequal distributions that allow the greater advantages to some to outweigh disadvantages to others *and are also to everyone's advantage in comparison to equality* are practically unfeasible; it does not rule out the possibility of 'aggregating' distributions (as I can call distributions that allow advantages to some to outweigh disadvantages to others) altogether. The simple assumption of the first model would still admit the practical feasibility of aggregating distributions which were *not* to everyone's advantage in comparison to equality. In support of reading **Passage 2g (*J as F 2*)** as allowing the practical feasibility of aggregating distributions which are not to everyone's advantage in comparison with equality, I can point out that not only does the passage do nothing to imply that such distributions are unfeasible, but that it would contradict Sentence [6]'s assertion that the 'restriction' expressed in **Passage 2g (*J as F 2*)**[1] is 'a restriction of consequence.'[2] It would be a very *inconsequential* restriction that only forbade distributions which were practically unfeasible, in any case.

2.1.26 If aggregating distributions which are *not* to everyone's advantage in comparison to equality are allowed into the feasible set, but aggregating distributions which are to everybody's advantage in comparison to equality are excluded, then the feasible set might be something similar to that represented in the figure below

Fig 2 (iii)

Distribution	Non-agricultural workers	Agricultural workers
1	115	140
3	100	100
4	170	90

[1] p. 88.
[2] Rawls describes it as 'a' restriction though, as the argument of these few sections shows, whether it expresses just one restriction is questionable. The fact that he suggested it did provides another ground for supposing that he was committed to the simple assumption of the first model.

2.1.27 For the purpose of my argument the exact figures in Distribution 4 do not matter. What does matter, is that Distribution 4 represents a feasible aggregating distribution, and *any feasible aggregating distribution would have the effect of placing some people below the equal distribution, 3, just as Distribution 4 does*. There would, then, be a very powerful line of argument in favour of requirement (c) from the point of view of Justice as Reciprocity: any aggregating distribution – that is, any distribution that justifies inequalities on the grounds that the disadvantages of those in one position are outweighed by the greater advantages of those in another position, such as Distribution 4 on either interpretation (i) or (ii) of advantage or disadvantage – would violate **the mutual advantage condition** by not being to everyone's advantage in comparison to the relevant situation of equal liberty.

2.1.28 So the simple assumption of the first model would have rendered the question of how to interpret **Passage 2g (*J as F 2*)** unimportant from a practical point of view. On either interpretations (i) or (ii) of 'advantage' and 'disadvantage', justifying inequalities on the grounds that the 'disadvantages' of those in one position are outweighed by the 'greater advantages' of those in another position would violate **the mutual advantage condition**. So 'aggregation' would violate the mutual advantage condition however 'aggregation' might be interpreted.

2 Justice as fairness in the first form of the model

2.2.1 In this section I argue first, that Justice as Fairness in the first model contained a much stronger line of argument against the principle of utility and for the two principles of justice than his subsequent model would and secondly, that this line of argument seems to obviate the need for Justice as Fairness to play a role in Rawls's theory of Justice as Reciprocity altogether.

2.2.2 As there were important difference between Rawls's principles of the first model and those of *Theory*, so too, there were some very important differences between Rawls's theory of Justice as Fairness in his first model and in *Theory*.

2.2.3 In the equivalent of *Theory*'s original position the decision making parties are, as they would later be in *Theory,* conceived of as 'mutually self-interested.'[1] However, they are not, as they would later be in *Theory*, to be imagined as making their choices from behind a 'veil of ignorance', but instead are fully aware of their position, and of how the practices of the 'hypothetical society'[2] they are members of affect them. The practices of the hypothetical society are not generally assumed to be 'just', i.e. compliant with the principles of justice, though some may happen to be. As Rawls described the decision problem facing the parties

Passage 2h (*J as F 2)*

> Since these persons are conceived as engaging in their common practices, which are already established, there is no question of our supposing them to come together to deliberate as to how they will set these practices up for the first time. Yet we can imagine that from time to time they discuss with one another whether any of them has a legitimate complaint against their established institutions.[3]

2.2.4 The first model's equivalent of *Theory*'s 'veil of ignorance' is Rawls's supposition that

Passage 2i (*J as F 2)*

> [t]hey [the parties in the first original position] each understand further that the principles proposed and acknowledged on this occasion are binding on future occasions. Thus each will be wary of proposing a principle which would give him a peculiar advantage, in his present circumstances, supposing it to be accepted. Each person knows that he will be bound by it in future circumstances the peculiarities of which cannot be known, and which might well be such that the principle is then to his disadvantage. The idea is that everyone should be required to make *in advance* a firm commitment, which others also may reasonably be expected to make, and that no one be given the opportunity to tailor the canons of a legitimate complaint to fit his own special condition, and then discard them when they

[1] Rawls 1958 p.168
[2] Rawls 1957 p.656
[3] Rawls 1958 p 171

no longer suit his purpose.[1]

2.2.5 The first model, then, relied on the supposition that the parties' ignorance of their future circumstances, coupled with the requirement that the principles chosen must bind them in such circumstances, would be strong enough to produce principles that could be unanimously agreed to by people aware of their present circumstances. On initial inspection, it doesn't appear to be strong enough for that task. The principal source of difficulty is, as Wolff puts it, 'the impossibility of achieving unanimity among a group of players who, in a manner of speaking, know too much about themselves and their fellow-players.'[2] This difficulty provides at least part of the explanation for why Rawls introduces the veil of ignorance in subsequent models.

2.2.6 For the sake of my argument, let us suppose that they would choose the two principles of justice and consider what implications that choice would have for Rawls's theory of reciprocity as it was in the first model. First we need to be clear about what Rawls meant by 'justice as fairness' in his first model.

2.2.7 In 'Justice as Fairness' (1), Rawls wrote

Passage 2j (*J as F 1*)

> These remarks [regarding the rationality of choosing the two principles in the hypothetical society] are not, of course, offered as a proof that persons so circumstanced would settle upon the two principles, but only to show that the principles of justice could have such a background; [1] *and so can be viewed as those principles which mutually self-interested and rational persons, when similarly situated and required to make in advance a firm commitment could acknowledge as restrictions governing the assignment of rights and duties in their common practices, and thereby accept as limiting their rights against one another.*
>
> 3. That the principles of justice can be regarded in this way is an important fact about them. It brings out the idea that fundamental to justice is the concept of fairness which relates to right dealing between persons who are cooperating with or competing against one another, as when one speaks

[1] Rawls 1958 pp 171-172
[2] Wolff 1977 p 51

> of fair games, fair competition, and fair bargains. The question of fairness arises when free persons, who have no authority over one another, are engaging in a joint activity and amongst themselves settling or acknowledging the rules which define it and which determine the respective shares in its benefits and burdens. A practice will strike the parties as fair if none feels that, by participating in it, he, or any of the others, is taken advantage of, or forced to give in to claims which he does not regard as legitimate. [2] *A practice is just, then, when it satisfies the principles which those who participate in it could propose to one another for mutual acceptance under the aforementioned circumstances.* Persons engaged in a just, or fair, practice can face one another honestly, and support their respective positions, should they appear questionable, by reference to principles which it is reasonable to expect each other to accept. *[3] It is this notion of the possibility of mutual acknowledgement which makes the concept of fairness fundamental to justice.* Only if such acknowledgement is possible, can there be true community between persons in their common practices; otherwise their relations will appear to them as founded to some extent on force and violence.[1] [My italics and numbering of the sentences]

2.2.8 The three italicized sentences encapsulate Rawls's theory of Justice as Fairness as it was in his first essay of the same name.[2] The definition Rawls gives of Justice as Fairness in *Theory* is essentially the same, apart from the removal of the reference to practices.[3] But the reasons for the parties choosing the two principles are very different to those they would become in *Theory.* In 'Justice as Fairness' (1&2) the parties choose the principles because they are the only principles that ensure that all would do as well, or better, than they would under an equal economic distribution, which was also a distribution of equal liberty. So in the first form of the model, Justice as Fairness held that it was both a necessary and sufficient condition for the parties' choice of principles, that the principles guaranteed that all would be at least as well off as they would be in a situation of equal liberty.

2.2.9 Now this raises a puzzling question regarding Rawls's first form of the model, which is why did he feel any need to support the two principle of justice with the

[1] Rawls 1957 p.657

[2] The wording in 'Justice as Fairness' (2) is exactly the same apart from the insertion of 'or fair' after 'is just' in the equivalent of the second italicized sentence. Rawls 1958 p.178

[3] It reads '[t]he original position is, one might say, the appropriate initial status quo, and thus the fundamental agreements reached in it are fair. This explains the propriety of the name "justice as fairness": it conveys the idea that the principles of justice are agreed to in an initial situation that is fair.'

contractualist argument of 'Justice as Fairness'? I can only put forward a speculative answer to that question, but in so doing I can also put forward the argument I promised earlier to the effect that Justice as Fairness should give way to the requirement that all people do at least as well as they would in the relevant situation of equal liberty in the event of a clash between meeting that requirement, and meeting the contractualist requirement of Justice as Fairness.

2.2.10 At the heart of Rawls's conception of Justice as Reciprocity there lies, I believe, the idea that cooperating parties in a venture for mutual advantage should be treated as if they were free and equal. He said this in so many words in the selection from *Justice as Fairness: A Restatement* that I cited as **Passage P3.** To repeat these: the historical conception of society that Rawls embraces as the primary alternative to utilitarianism is 'the idea of society as a fair system of social cooperation between citizens regarded as free and equal'. The conception of justice that I have attributed to Rawls, Justice as Reciprocity, holds that rules, or institutions, are 'just' insofar as they are appropriate to that conception of society. In this case, Justice as Reciprocity, should, given the assumptions of Rawls's first model, be able to mandate the two principles of justice directly without any need for the 'middle man' of Justice as Fairness. Justice as Reciprocity would seem, at least on the face of it, to require that all cooperating members of society do at least as well as they would in the relevant situation of equal liberty. According to the assumptions of the first model, as I have been at pains to emphasize in these last two sections, the principles of justice would meet this requirement, *and they are the only principles*, that could meet this requirement. In terms of Fig. 1(i) in Chapter 1, Justice as Reciprocity could proceed via the Justice as Fairness bypass 1 straight to the two principles of justice. This implication of Rawls's first model explains why I did not feel the need to enter into the argument of Justice as Fairness, that the parties would choose the principles in the original position, in too much detail. If Justice as Fairness did not select the two principles of justice then it would appear to be unfit for the purpose of constructing principles of Justice as Reciprocity. I return to this line of argument in Chapter 3, where I deploy it to dismiss Rawls's theory of Justice as Fairness as it was in *Theory.*

2.2.11 My speculation as to why Rawls felt the need to deploy the contractualist argument

of Justice as Fairness is that contractualism is, as Rawls repeatedly emphasizes, the great historical alternative to utilitarianism, and one that seems closely allied to Rawls's conception of society.

2.2.12 It is worth remarking here that Rawls's use of a *hypothetical* contract seems to me to defeat at least one of the main purposes behind the invocation of a contract in traditional social contract theory. Contractualist theorists such as Locke and Hobbes depended heavily on the idea that the social contract was binding because breaking it would be breaking an *actual* contract, and breach of contract was unjust. Rawls explicitly repudiates the idea that justice is akin to an actual contractual obligation throughout the various developments of his theory.[1]

3 The duty of fair play

2.3.1 What was to become the principle of fairness in *Theory* made its debut as the duty of fair play in 'Justice as Fairness' (1). It plays a comparatively minor role in *Theory*, where Rawls holds it to obligate only persons who have voluntarily accepted offices in a well-ordered just society, and not citizens generally. Indeed, Rawls remarks in *Theory* that '[t]here is, I believe, no political obligation, strictly speaking, for citizens generally.'[2] But I believe the duty of fair play was central to Rawls's first model and was intended to apply to

[1] Ronald Dworkin questioned the point of Rawls's hypothetical contract in his essay, 'Justice and Rights' (1978 p. 151), remarking that '[a] hypothetical contract is not simply a pale form of an actual contract; it is no contract at all.' Dworkin (who was commenting on Rawls's third model of *Theory*) went on to suggest that the device of choice in the original position from behind a veil of ignorance would be more usefully understood as a device for testing which principles would best fulfilled the obligation of political institutions to treat the individuals they govern equally. However, in 'Justice as Fairness: Political not Metaphysical.' (p. 236), Rawls rejects Dworkin's suggestion on the grounds that it is too 'narrow' an interpretation of justice as fairness. One of the reasons Rawls gives for regarding Dworkin's interpretation as too narrow is that justice as fairness is rooted in the 'fundamental and intuitive idea of society as a fair system of cooperation.' But if that is the fundamental idea, if the two principles of justice are the only principles of justice that could fulfil the mutual advantage condition, then the need for justice as fairness is questionable. Rawls was, of course, defending the third model's version of justice as fairness which is very different to the first. However, in Chapter 3, I go on to argue that the third model of justice as fairness is unsuited to the fundamental idea of society as a fair system of cooperation precisely because it doesn't guarantee principles that would meet the mutual advantage condition of Justice as Reciprocity.

[2] Rawls *T of J Rev* 1999 p. 98

all citizens. It sets out the motive to be just, and, I shall maintain, exactly corresponds to Gibbard's defining features of Justice as Reciprocity: that is, those who had a duty to act in accordance with the duty of play, as it should be interpreted in Rawls's first model are those who would have a duty of reciprocity according to Gibbard's definition of Justice as Reciprocity. (§§1.4.1 -1.4.2). I shall also show that only Rawls's two principles of justice could meet the duty of fair play in Rawls's first model; the classical principle of utility could not do so.

2.3.2 I should explain the significance of this result for the conclusion of my thesis as a whole. When I first started this thesis I nursed the ambition to show that the principle of classical utility was, contra Rawls, the principle that was uniquely suited to Justice as Reciprocity and that the motive for acting in accordance with the principle of classical utility would be the duty of fair play, i.e. a duty of reciprocity, rather than any putative duty of benevolence. I have curtailed that ambition to the more modest one of showing that the classical principle of utility is reconcilable with the conception of Justice as Reciprocity. The reason for this curtailment is that I have not solved the problem, and do not believe that anyone else has either, of what the fairness condition might entail.[1]

2.3.3 The Rawls of the first model thought he had solved the problem and believed that the fairness condition effectively amounted to the same thing as the mutual advantage condition. He did not make this belief particularly clear but it can be discerned by a close reading of what he wrote, which I provide below. This might seem surprising, as the fairness condition and the mutual advantage condition are *conceptually* distinct requirements and were presented as such in Chapter 1; the reason they were not *practically* distinct, I shall explain, was due to Rawls's theory of Justice as Fairness.

2.3.4 But Rawls had not solved the problem, and the repercussion of this failure was to leave open the question of what the fairness condition might entail, which in turn left open the question of whether the duty of fair play could as equally well apply to a society that is well-ordered by the principle of classical utility as to one that is well-ordered by the two

[1] I confess to still nursing the same ambition, though it is beyond the remit of this thesis to fulfil it.

principles of justice. In Chapter 4 I argue that other things indeed being equal, the ideal legislator should feel free to choose the classical principle of utility if that is where he or she felt that choice would do the most good. Which is also to say that, although Justice as Reciprocity and Justice as Benevolence are indeed two distinct *conceptions* of justice, their *practical* recommendations in terms of a governing principle of distributive justice could coincide.

2.3.5 Rawls introduces the duty of fair play in the course of providing his account of why people should be bound by a practice that meets the principles of justice, in a passage just following **Passage 2j (*J as F 1*)**

Passage 2k (*J as F 1)*

> [1]Now if the participants in a practice accept its rules as fair, and so have no complaint to lodge against it, there arises a prima facie duty (and a corresponding prima facie right) of the parties to each other to act in accordance with the practice when it falls upon them to comply. [2] When any number of persons engage in a practice, or conduct a joint undertaking, according to rules, *and thus restrict their liberty*, those who have submitted to these restrictions when required have a right to a similar acquiescence on the part of those who have benefited by their submission. [3] These conditions will, of course, obtain if a practice is *correctly* acknowledged to be fair, for in this case, all who participate in it will benefit from it. [my italics and numbering of the sentences][1]

2.3.6 This statement of the duty of fair play, if taken out of context, might appear to be compatible with a very wide range of principles of distribution. If all that is required for a practice to qualify as fair is, as sentence [1] implies, that its practitioners accept it as fair, then the 'fairness' of any practice would seem to be purely subjective, dependent on the opinions of the practitioners. If all accepted the two principles of justice as fair, then the two principles of justice would be fair. But if all accepted the principle of utility as fair then the principle of utility would be fair.

2.3.7 But in the context of the account of Rawls's two principles of justice and Justice as

[1] Rawls 1957 p.665

Fairness as they were in 'Justice as Fairness', just given, the passage should be read as restricting the duty of fair play to *only* those practices that are in accordance with Rawls's principles of justice. This reading is justified by paying close attention to sentences [2] and [3]. The 'conditions' of sentence [3] are those conditions a practice needs to meet in order for people to have a prima facie duty to act in accordance with that practice in their capacity as beneficiaries of that practice, and a prima facie right to expect a similar acquiescence from other beneficiaries of that practice in their capacity as 'contributors' to the benefits of the practice. The participants in the practice qualify as 'contributors', due to the fact that their restriction of their liberty has, Rawls presumes, increased the size of the benefits available to all participants in the practice.

2.3.8 Now sentence [3] might at first appear to be strange. It asserts that *if* a practice is correctly acknowledged to be fair *then* those who participate in it will benefit from it. That is, it asserts that if the antecedent of the conditional (that participants in a practice correctly regard the practice as fair) obtains, then the consequent (that everyone benefits from the practice) obtains as well. The reason this assertion might seem odd is that the idea of an inference from someone's accepting a practice as fair to someone's benefiting from it is highly questionable. To use an example that Rawls would have approved of at the time of writing 'Justice as Fairness': an indoctrinated slave might perceive the practice of slavery to be fair, but a slave surely does not benefit from his slavery.[1] However, we should pay particular attention to Rawls's qualification that the practice has to be 'correctly' acknowledged as fair, and also to his theory of Justice as Fairness. Rawls has argued that *rational* persons would only choose the two principles of justice to govern their practices, and this should be read as what he means by 'correctly' in this passage. So we can infer that those practices that are correctly acknowledged as fair *would* benefit the participants in the practice. This is because, as shown in the preceding section (§2.2.8), the parties in the first original position would choose the two principles of justice precisely because they did

[1] I write 'at the time' advisedly. I shall presently quote a passage from 'Justice as Fairness' which demonstrates that Rawls's first model did not construe the practice of slavery to be beneficial to the slaves. And as already shown, in **Passage 1n (*DJ 1967)*** (p. 63), he took the same position in the significantly revised model of 'Distributive Justice' (1967). But, as evidenced by **Passages 1r and 1t (*Theory Rev)*** by the time of *Theory*, he was committed to a definition of advantage with respect to the baseline of a Hobbesian state of nature, according to which a system of slavery might prove to be to everyone's advantage.

offer an improvement on the relevant situation of equal liberty.[1]

2.3.9 It should be noted that it would also be the case that all those practices that benefited all those who participated in them would also be fair. This is guaranteed by the simple assumption of the first model that there could be at most only one distribution which was to everyone's advantage in comparison with the same practice under equality, together with the assumption that the parties in the first original position would choose principles to govern their practices that guaranteed that everyone fared at least as well as under the same practice under equal liberty. So if the antecedent of the conditional (that everyone benefits from the practice) obtained, then the consequent (that participants in a practice correctly regard the practice as fair) would follow.

2.3.10 The relationship between 'fairness' and 'benefit', then, in the first model was one of equivalence. A practice or venture would be fair if, and only if, it would also be beneficial to its participants. This is also to say that a practice or venture would be beneficial to its participants if, and only if, it were fair. This completes my demonstration that the two conceptually distinct requirements of Justice as Reciprocity did not lead to practically distinct requirements given the assumptions of the first model.

2.3.11 This relationship between 'benefit' and 'fairness' in the first model carried with it two important implications for my overall argument. First, the duty of fair play would only bind people to practices or ventures that were governed by the two principles of justice. And secondly, the two principles of justice would be the only conception of justice that would fit what I referred to in Chapter 1 as Rawls's conception of Justice as Reciprocity. This took Justice as Reciprocity to be defined by three features: firstly, justice would require constraint on the part of those who are required to act justly. Secondly, the fairness condition, which stipulates that a necessary condition that must be met for any individual

[1] It might seem that the two principles shouldn't be interpreted as offering an improvement on the relevant situation of equal liberty as if the second principle hasn't come into play, the first principle just provides participants with equal liberty. But it should be noted that the first principle refers to 'the most extensive liberty compatible with a like liberty for all'. The relevant situation of equal liberty in comparison to which the combined principles offer an improvement can, then, be interpreted as a less extensive liberty than that provided by the first principle of justice.

participating in a cooperative venture to have a duty of reciprocity, is that the rules governing the cooperative venture are fair. And thirdly, the mutual advantage condition which stipulates that a second condition that must be met for any individual participating in a cooperative venture is that the venture affords them some prospect of 'advantage'. The utilitarian conception of justice, whereby practices of joint ventures are governed by the principle of utility, would, according to Rawls's first model, meet neither the fairness condition nor the advantage condition. So no one could be bound by the duty of fair play to utilitarian practices. I return to reconsider what the relationship of the utilitarian conceptions of justice would be to Justice as Reciprocity under the assumptions of *Theory* at the end of this Chapter.

4 Justice as Reciprocity v Justice as Benevolence in Rawls's first model

2.4.1 Rawls first used the term 'Justice as Reciprocity' in his essay 'Justice as Fairness' (1). In this section I analyse what he meant by the term and contrast it with his conception of Justice as Benevolence, as it was in the first model. Having described the ideas, I can then demonstrate how, given the assumptions of Rawls's first model, the expectation that the two competing conceptions of justice would yield similar policy prescriptions looked at least tenuous. This will enable me to show, in the Concluding Remarks of the next section, how, given the revised assumptions of *Theory* the expectation that they would yield similar policy prescriptions didn't seem nearly so tenuous.

2.4.2 In 'Justice as Fairness'(1), Rawls provides a summary of the first model of his theory which he then connects directly to what he calls 'conditions of reciprocity and community'. The paragraph containing this summary and the paragraph following it, defining the condition of reciprocity, are very useful. They illustrate just how closely connected the different elements of Rawls's theory, Justice as Fairness, the two principles of justice and the duty of fair play, were in the first model.[1] I repeat those paragraphs below

Passage 21 (*J as F 1*)

[1] This is the 'neatness of fit' referred to earlier (§2.1.23)

> [1] The conception at which we have arrived, then, is that the principles of justice may be thought of as arising once the constraints of having a morality are imposed upon rational and mutually self-interested parties who are related and situated in a special way. [2] A practice is just if it is in accordance with the principles which all who participate in it might reasonably be expected to propose or to acknowledge before one another when they are similarly circumstanced and required to make a firm commitment in advance; and thus when it meets standards which the parties could accept as fair should occasion arise for them to debate its merits. [3] Once persons knowingly engage in a practice which they acknowledge to be fair and accept the benefits of doing so, they are bound by the duty of fair play which implies a limitation on self-interest in particular cases.
>
> [4] Now if a claim fails to meet this conception of justice there is no moral value in granting it, since it violates the conditions of reciprocity and community amongst persons: he who presses it, not being willing to acknowledge it when pressed by another, has no grounds for complaint when it is denied; whereas him against whom it is pressed can complain. [5] As it cannot mutually be acknowledged, it is a resort to coercion: granting the claim is only possible if one party can compel what the other will not admit. [6] Thus in deciding on the justice of a practice it is not enough to ascertain that it answers to wants and interests in the fullest and most effective manner. [7] For if any of these be such that they conflict with justice, they should not be counted; their satisfaction is no reason for having a practice. [8] It makes no sense to concede claims the denial of which can be objected to. [9] It would be irrelevant to say, even if true, that it resulted in the greatest satisfaction of desire.[1]

2.4.3 The usefulness of this passage lies in its exposure firstly; of just which claims people would have the right to press, with others having an obligation to grant and secondly; which claims they wouldn't have the right to press and others would have no duty to grant, according to Justice as Reciprocity in Rawls's first model. 'The principles of justice' referred to in sentence [1] are, of course, Rawls's two principles of justice. Sentence [2] defines those practices as 'just' as those practices which could be acknowledged by those in 'the original position'. This also, it is reasonable to suppose, is the same conception of justice at work in sentences [7] and [8]. Only Rawls's principles of justice could be acknowledged in the 'original position' and, on my interpretation given above (§§2.1.11 - 2.1.21), they would be acknowledged because they are the only principles that would ensure that everyone did as well, or better, than under equality.

[1] Rawls 1957 p.660

Sentence [3], in line with my interpretation of the duty of fair play given above, (§§2.3.1 - 2.3.7) can be read as asserting that since only those practices that are in accordance with the two principles of justice could qualify as fair, only those practices could bind people by the duty of fair play.

2.4.4 In anticipation of the case that I shall put in the next section for the principle of classical utility's compatibility with Justice as Reciprocity I shall point out here that at the root of Rawls's conception of justice is the assumption that only the two principles of justice would ensure they would all do as well or better than in the relevant situation of equal liberty. This is why they are chosen in the first original position, making them 'fair'. And the duty of fair play, as shown in Section 3, only obligates people to principles which are fair.

2.4.5 The next question to be addressed is what Rawls meant by 'the conditions of reciprocity' in sentence [4]. This turns out to be a comparatively simple question to answer. Two pages before **Passage 21 (*J as F 1*)** Rawls wrote

Passage 2m (*J as F 1*)

> It is this notion of the possibility of mutual acknowledgment which makes the concept of fairness fundamental to justice. Only if such acknowledgment is possible, can there be true community between persons in their common practices; otherwise their relations will appear to them as founded to some extent on force and violence.[1]

2.4.6 Rawls nowhere distinguishes between conditions of community and reciprocity, so it is reasonable to construe them as being the same. In that case, conditions of reciprocity as those conditions which would be met by claims that conform to principles that could be mutually acknowledged in the first original position. The condition of reciprocity, then, appears to be met by the same practices as would meet Justice as Fairness. Rawls's point in invoking the term 'reciprocity' appears to be to emphasize the consensual nature of persons acting in accordance with these principles as opposed to those which violate those

[1] Rawls 1957 p. 668.

conditions which are founded on force', 'violence' or 'coercion', to use the terms evoked in **Passage 21 (*J as F 1*)** and **Passage 2m (*J as F 1*).**

2.4.7 This interpretation of Rawls's use of the term 'reciprocity' in Rawls's first model is supported by the definition of a principle of reciprocity Rawls put in a later essay, 'Justice as Reciprocity' (1971), which was essentially a rewrite of 'Justice as Fairness' (2) for an edited collection.[1] I put this definition here, but will refer back to it in Chapter 3 as a possible candidate for a mysterious principle of reciprocity that Rawls refers to in *Theory.*

Passage 2n (*J as R 1971*): the principle of reciprocity

> The principle of reciprocity requires of a practice that it satisfy those principles which the persons who participate in it could reasonably propose for mutual acceptance under the circumstances and conditions of the hypothetical contract. Persons engaged in a practice meeting this principle can then face one another openly and support their respective positions, should they appear questionable, by reference to principles which it is reasonable to expect each to accept. A practice will strike the parties as conforming to the notion of reciprocity if none feels that, by participating in it, he or any of the others are taken advantage of or forced to give in to claims which they do not accept as legitimate.[2]

2.4.8 My interpretation of Rawls's 'conditions of reciprocity' as being met by the same condition as Justice as Fairness can be confirmed by considering that if 'Justice as Fairness' were substituted for 'the principle of reciprocity' in **Passage 2n (*J as R 1971*)** it would still make perfect sense. In this passage, as in the previous two, the term appears to be evoked to emphasize the willing nature of reciprocal behaviour, as opposed to being 'taken advantage of' or 'forced'.

2.4.9 In fact, Rawls substituted 'justice as fairness' for 'justice as reciprocity' himself. **Passage 2o (*J as F 1)*** below contains Rawls's first use of the phrase 'justice as reciprocity'; there is an equivalent passage in 'Justice as Fairness' (2) which is essentially

[1] Originally published in Samuel Gorovitz, ed., *Utilitarianism: John Stuart Mill: With Simple Essays,* pp. 242–268. New York: Bobbs-Merrill, 1971.
[2] Rawls 1971 p.208

the same except for the substitution of 'justice as fairness' for 'justice as reciprocity'.[1]

2.4.10 Turning now to consider which claims would violate the condition of reciprocity; these are those which might be pressed which aren't in accordance with principles that could be mutually acknowledged. We have already seen that Rawls regarded them as being founded on 'force', 'violence' or 'coercion.' But the fact that they are not in accordance with principles that are mutually acknowledged does not mean that they could not be given some other principled justification, however. It is obvious that Rawls has the principle of utility in mind when he writes of claims that would violate the condition of reciprocity 'it would be irrelevant to say…that it resulted in the greatest satisfaction of desire.'[2]

2.4.11 So in summary: Rawls's conception of Justice as Reciprocity in the first model put the emphasis on the fact that principles that conform to the conception must be mutually acknowledged, and that is how his use of the term 'reciprocity' should be understood.[3] However, the three main elements of his theory in the first model were so inextricably intertwined that any principles that satisfied Justice as Reciprocity would be the only principles that would satisfy the other elements of his theory: the only principles that could be mutually acknowledged would also be the only principles that could meet the requirement of the duty of fair play. Similarly, the only principles that advantage all with respect to the relevant position of equality would be the only principles that could be selected by Justice as Fairness. Furthermore, the two principles of justice would be the only principles that would meet the conditions of any of these elements of Rawls's conception of justice in the first model. And, as I already demonstrated in Section 3 on 'the duty of fair play' (§2.3.8), the two principles of justice would be the only conception of justice capable of meeting the requirements of Gibbard's definition of Justice as Reciprocity set out in Chapter 1 (§§1.4.1 - 1.4.2).

2.4.12 So I turn now to Rawls's conception of Justice as Benevolence as it was in the first

[1] See footnote 43 below
[2] in the words of **Passage 21 (*J as F 1*)'s Sentence [9]**
[3] I haven't analysed Rawls's use of the actual term 'Justice as Reciprocity' in 'Justice as Fairness' (1). This occurs in **Passage 2o (*J as F 1)*** below and is entirely in conformity with the conception outlined here.

model. He starts by defining it in opposition to his conception of Justice as Reciprocity in the paragraph immediately following **Passage 21 (*J as F 1*)**. In fact, this paragraph is the one cited as **P3** in my Preface, which I repeat here for convenience

Passage P3 ***(J as F 1957)***

> This conception of justice (i.e. the conception of justice as reciprocity) differs from that of the stricter form of utilitarianism (Bentham and Sidgwick), and its counterpart in welfare economics, which assimilates justice to benevolence and the latter in turn to the most efficient design of institutions to promote the general welfare.[1]

2.4.13 The point to be noted, as was already noted in my Preface, is that Rawls's pictures of the rival conceptions of society were remarkably similar in his first essay 'Justice as Fairness' in 1957 to what they would be in *Justice as Fairness: A Restatement* (2001)

2.4.14 Rawls considers that Justice as Benevolence might recommend the two principles of justice as a practical means to maximizing utility, but points out that, if so, the two ideas behind the rival conceptions of justice would still be very different.

Passage 2o (*J as F 1)*

> But even if such restrictions [i.e. the assumption that all members of society have similar utility functions and the assumption of the law of diminishing marginal utility] are built into the utility function, and have, in practice, much the same result as the application of the principles of justice (and appear, perhaps, to be ways of expressing these principles in the language of mathematics and psychology), the fundamental idea is very different from the conception of justice as reciprocity.[2] Justice is interpreted as the contingent result of a higher order administrative decision whose form is

[1] Rawls 1957 p.660

[2] The equivalent of this phrase and surrounding text in J as F (2) is '(and appear, perhaps, to be ways of expressing these principles in the language of mathematics and psychology), the fundamental idea is very different from the conception of justice as fairness.[2] For one thing, that the principles of justice should be accepted is interpreted as the contingent result of a higher order administrative decision. The form of this decision is regarded as being similar to that of an entrepreneur…' (Rawls 1958, p.185) A comparison of the two selections should justify my claim above that they are 'essentially the same' (§2.4.8).

> similar to that of an entrepreneur deciding how much to produce of this or that commodity in view of its marginal revenue, or to that of someone distributing goods to needy persons according to the relative urgency of their wants. [my underlining][1]

2.4.15 Next, Rawls introduces the figure of the ideal legislator for the first time

Passage 2p (*J as F 1)*

> The individuals receiving the benefits are not thought of as related in any way: they represent so many different directions in which limited resources may be allocated. Preferences and interests are taken as given; and their satisfaction has value irrespective of the relations between persons which they represent and the claims which parties are prepared to make on one another. This value is properly taken into account by the (ideal) legislator who is conceived as adjusting the rules of the system from the center so as to maximize the present capitalized value of the social utility function.[2]

2.4.16 This is enough textual evidence from 'Justice as Fairness' (1), to support a couple of important points about Rawls's conception of Justice as Benevolence in his first model, and to contrast it with his conception of Justice as Benevolence in *Theory*.

2.4.17 The first point to note, which is very important to the overall argument of this thesis, is that Justice as Benevolence was conceived of as starting at a lower level of construction to how it would be conceived in *Theory*. There was no mention of the impartial observer, or the utilitarian conception of society, nor would there be for several essays to come.[3] In terms of Fig. 1(i) Justice as Benevolence started at the lower level of construction, referred to as 'the utilitarian theory of justice.' The reason this is important is that utilitarianism can very easily be conceived of starting at this lower level, in which case strong lines of arguments directed against the higher level of construction may prove to be misleading distractions. I shall argue that 'the separateness of persons' objection to 'the

[1] Rawls 1957 p. 661
[2] Rawls 1957 p.662
[3] The utilitarian conception of society would make its first appearance in 'Constitutional Liberty and the Concept of Justice' (1963). The impartial spectator wouldn't make his debut until 'Distributive Justice: Some Addenda' (1968)

utilitarian conception of society'[1] is a misleading distraction in Chapter 4.

2.4.18 The second point to note is that although Rawls conceded that the two conceptions of justice may converge on the same practical result of recommending the two principles of justice, their convergence would be precarious. **Passage 2o (*J as F 1)*** described the 'practical result' of utilitarianism recommending the principles of justice as 'the contingent result of a higher order administrative decision'.[2] Rawls's point here is that utilitarianism would only support the two principles of justice if 'the ideal legislator', or other higher order utilitarian administrative device, estimated that doing so would maximize utility. If the ideal legislator estimated the two principles of justice wouldn't maximize utility, he or she wouldn't prescribe them. This would be in contrast to the conception of Justice as Reciprocity which would prescribe them on the grounds that they would be the only principles that could be mutually acknowledged regardless of whether they maximized utility.

2.4.19 The two conceptions of justice, then, seem quite likely to lead to different policy prescriptions. Even if the utilitarian ideal legislator would select the two principles of justice as rules of thumb, he would not require them to be as strictly observed as would the conception of Justice as Reciprocity. As I remarked in Chapter 1, even as strong a champion of individual liberty as John Stuart Mill allowed that his liberty principle should give way to considerations of expediency on occasions where the calculus of utility called on it to do so (§1.5.5).

5 What if Rawls were not committed to the 'simple assumption of the first

[1] By which I mean Rawls's higher level of account of how utilitarianism might best be motivated, referred to in footnote 46 above.

[2] Rawls later questions utilitarianism's reliance on the empirical assumptions of similar utility functions and diminishing marginal utility to deliver the same practical results as Justice as Reciprocity, claiming that 'the various restrictions on the utility function needed to get this result are borrowed from the conception of justice as fairness. The notion that individuals have similar utility functions, for example, is really the first principle of justice under the guise of a psychological law. It is assumed not in the manner of an empirical hypothesis concerning actual desires and interests, but from sensing what must be laid down if justice is not to be violated**.'** (Rawls 1957 p.662) This strikes me as a display of excessive enthusiasm for the two principles of justice on Rawls's part. The notion that individuals have similar utility functions seems a reasonable empirical hypothesis.

model'?

2.5.1 Thus far I have proceeded on the supposition that Rawls was committed to the simple assumption of the first model, because it is my belief that he was. But a powerful argument against this supposition is that it suggests that he was a bit naïve: surely he should have been alert to the possibility of there being more than one unequal distribution that is Pareto-preferred to equality? I have seen no evidence to suggest that Rawls was alert to that possibility at the time, and provided some grounds for supposing that he wasn't. Other commentators, however, have helpfully attempted to 'fill in the gaps' in Rawls's work and argue that the difference principle is the most 'egalitarian' compromise on actual equality. Thoroughness recommends the consideration of at least one of these alternatives; after all, there may be an argument that would have served Rawls's purposes equally well, whether he considered it or not.

2.5.2 So I here consider an argument offered by Brian Barry in *Justice as Impartiality,*[1] where Barry's reference to Broome's example suggests that he considered his new justification would work for examples such as Broome's. I argue below that it doesn't.

2.5.3 The purpose of Barry's argument of *Justice as Impartiality* was to side with Rawls in defence of the difference principle against utilitarianism. He ends up concluding that 'utilitarianism… fails as a theory of justice because it is liable to place unfair burdens on some people' and 'the inequalities in well-being that it [i.e. utilitarianism] can give rise to cannot be justified to those who are at the losing end of them.'[2]I can set aside the larger questions of whether utilitarianism is unfair, or can be justified to those at its losing end, to focus on the question in hand, which is whether Distribution 2 in Fig 2 (i) is unfair or can be justified to those at its losing end. This is the question in hand because the aim is to establish whether the difference principle is the most egalitarian interpretation of Rawls's second principle of justice in his first model. We are operating on the assumption that the requirement of Rawls's first principle of justice, that guarantees the greatest liberty compatible for all, has been met (it is represented by Distribution 3). So the more unequal

[1] Barry makes reference to Broome's example in Barry 1995 p. 66
[2] Barry 1995 p. 66

assumptions (that would endow people with less than 100) are not in the picture.

2.5.4 The easiest way for me to show that Distribution 2 could be justified to those who lose out under is by response to Barry's implicit argument that it couldn't, repeated in Passage 2r below

Passage 2r

> What then can we say to the actual worst-off people in a society where the difference principle is instantiated? We cannot say that *they* could not be made better off. But we can say that the currently worst-off people could be made better off only if some other people were made worse off than the currently worst-off people are now. The minimum level, in other words, would have to become lower. This necessarily follows from the stipulation that it is currently as high as it can be. Now, there is nothing to stop the members of the worst-off group saying 'Fine: make us better off at the expense of these others.' But this cannot be a *moral* demand. For the only moral basis for their being as well off as they are now is that the worst off people (whatever their identities) should be as well off as possible.[1]

2.5.5 For this argument to succeed in the face of Fig 2 (i) Barry would have to claim that the non-agricultural workers demand to make them better off than the agricultural workers could not be a moral demand. But there does seem to be a moral demand available to them, they could try saying "We're not just being greedy; it's fairer – there is less inequality in Distribution 2 than in Distribution 1; that seems intuitively obvious." And if aggregative considerations favoured Distribution 2, they could also say, "Plus there's more of us; more good will be done this way."[2]

2.5.6 Would Barry have a viable counter-argument to these moral justifications for the agricultural workers demands? The one implicit in the final sentence of Passage 2r is that 'the *only* moral basis' for people to have any more than they might otherwise have is 'that the worst off people… should be as well off as possible'. If it is true that the only possible

[1] Barry 1995 p.66
[2] These two lines of argument were suggested to me by Jonathan Wolff.

moral basis for anyone having any more than they might otherwise have is 'that the worst off people… should be as well off as possible' then it would be correct that the justifications the non-agricultural workers offer for their demand would not be 'moral' ones. But this purported counter-argument is clearly question begging: it does not establish the claim that there can be no other moral basis for people to have more than they might otherwise have; it simply assumes it.

2.5.7 Alternatively, someone might try to argue against allowing aggregative considerations into resolving the question of how the second principle of justice should be applied by some other 'egalitarian' criterion that would favour Distribution 2 on the grounds that it was fairer. Then Distribution 2 might be selected even in the case that there were many more agricultural workers than non-agricultural workers.

2.5.8 But at the end of the day, however the question of how best to interpret the second principle of justice in the face of the possibility of more than one unequal distribution that is Pareto-preferred to equality in the first model should be resolved, its resolution would make little difference to either Rawls's case for the two principles of justice being the principles most appropriate to Justice as Reciprocity on the one hand, or his case that utilitarianism is incompatible with Justice as Reciprocity on the other.

2.5.9 The reason that it would make little difference to his case that the two principles of justice were most suited to Justice as Reciprocity is that they would still be the only two principles that would meet the mutual advantage condition. This position in the first model, we can recall, is the assignment of equal rights that would provide people with the maximum level of net benefit compatible with the same level of net benefit for all. If, as Rawls maintained, Justice as Fairness lay at the heart of Justice as Reciprocity and the parties in the original position would choose the only principles that met the mutual advantage condition, they would choose the two principles of justice regardless of the subsidiary question of how to interpret the demands of the second principle of justice. And they would decisively reject utilitarianism for the reason that it would be liable to place some of them below the level of net benefit they would enjoy under the first principle of justice.

2.5.10 The point of §2.5.9 can be illustrated by adapting the example behind **Fig 2 (i)**. Suppose, for the sake of example, that Rufus T Firefly, the leader of Freedonia, had unlimited powers and chose to institute such a system of serfdom on the grounds that doing so would maximize utility because the agricultural workers would be more productive if tied to the land, only being allowed to travel outside the confines of the manor on high days and holidays. Rufus T Firefly in this example can be conceived as the 'ideal legislator' who is 'adjusting the rules of the system from the center so as to maximize the present capitalized value of the social utility function.'[1]

Fig 2 (v)

Distribution	Non-agricultural workers	Agricultural workers
1	115	140
2	120	110
3	100	100
4	150 (freemen and lords of the manor)	90 (serfs)

2.5.11 If there were enough non-agricultural workers then the principle of utility would select Distribution 4. But the agricultural workers would reject a principle that would be liable to give them less than they could be assured with the guarantee of equal liberty offered by the first principle alone. However, although they would presumably favour the interpretation of the second principle of justice that maximized the prospects of the worst off (resulting in Distribution 1) if it were unavailable, for some reason, they would still choose the two principles of justice because they would still be better off under the combination of the two principles than the first principle of justice alone.

2.5.12 So now it is possible to take full stock of the consequences of Rawls being, or not

[1]To quote **Passage 2p (*J as F 1)***

being, committed to the simple assumption of the first model. If he was so committed, then the two principles of justice would exclude any aggregation, whether it be understood according to interpretation (i) of Sentence 4 of **Passage 2g (*J as F 2*)**[1] or interpretation (ii). If he wasn't so committed, then there might be an argument that upheld the difference principle as the most egalitarian interpretation of the second principle of justice. In which case, the two principles of justice would again exclude any aggregation, understood under either interpretation. Alternatively, there might be another 'egalitarian' argument that would select one of the Pareto optimal unequal distributions on non-aggregative grounds as the best interpretation of the second principle of justice. Such an argument would also exclude aggregation on either interpretation. Finally, the second principle of justice might be interpreted as allowing for limited aggregation to resolve the question of which Pareto-optimal unequal distribution to select. In which case, the principle would exclude aggregation according to interpretation (ii) of the exclusion but not according to interpretation (i). The important point to note, however, is that in all these cases the *unlimited* aggregation that the principle of utility would allow would be excluded by the second principle of justice.

Concluding remarks

2.6.1 Here I take stock of the consequences of Rawls's abandonment of the assumptions of his first model and subsequent endorsement of a state of general egoism as the baseline from which mutual advantage should be measured. The way I shall do this is with reference to an applied example that Rawls conveniently supplied in 'Justice as Fairness' (1): that utilitarianism might recommend a system of slavery. He suggests this possibility in the following two passages, the first from 'Justice as Fairness' (1) and the second from 'Justice as Fairness' (2)

Passage 2s (*J as F 1*)

[1] p. 88

> [1] It [the utilitarian conception of justice] can lead one to argue against slavery on the grounds that the advantages to the slaveholder do not counter balance the disadvantages to the slave and to society at large burdened by a comparatively inefficient system of labor. [2] The conception of justice as fairness, when applied to the offices of slaveholder and slave, would forbid counting the advantages of the slaveholder at all. [3] These offices could not be founded on principles which could be mutually acknowledged, so the question whether the slaveholder's gains are great enough to counterbalance the losses to the slave and society cannot arise in the first place.[1]

Passage 2t (*J as F 2*)

> [1] Utilitarianism cannot account for the fact that slavery is always unjust, nor for the fact that it would be recognized as irrelevant in defeating the accusation of injustice for one person to say to another, engaged with him in a common practice and debating its merits, that nevertheless it allowed of *the greatest satisfaction of desire.* [2] The charge of injustice cannot be rebutted in this way. [3] If justice were derivative from a higher order executive efficiency, this would not be so. [2] [my italics and sentence numbering]

2.6.2 The essence of the argument contained between these passages is that the utilitarian conception of justice may or may not condemn slavery as unjust – the matter of whether it did so or not would depend on the contingent question of whether the advantages to the slaveholder would outweigh the disadvantages to the slave. However, suggests Rawls in Sentences [2] and [3] of **Passage 2s (*J as F 1*)**, Justice as Fairness would not hesitate in condemning slavery as unjust; slavery would be unjust because it would not be in accordance with principles that could be mutually acknowledged by the slaveholder and the slave.

2.6.3 Rawls's argument of these passages, incidentally, provides a fine example of a claim that would violate the conditions of reciprocity put forward in **Passage 21 (*J as F 1*).** That passage asserted that it would be irrelevant to say of a claim that was not put forward in accordance with principles that could be mutually acknowledged 'that it resulted in the greatest satisfaction of desire.' Such a claim would still violate the conditions of reciprocity and community. **Passage 2t (*J as F 2*) Sentence [1]** describes a potential

[1] Rawls 1957 p.661
[2] Rawls 1957 p.188

justification of slavery on the grounds that it allowed 'the greatest satisfaction of desire' as 'irrelevant' indicating that this is just the sort of claim that Rawls had in mind as being in violation of conditions of reciprocity and community.

2.6.4 But the crucial point to appreciate, when considering Rawls's argument, is that Justice as Fairness's unequivocal condemnation of slavery as unjust ultimately depends on the assumption that slavery would disadvantage the slave with respect to the relevant situation of equal liberty. This is why, to use the terms of **Passage 2s (*J as F 1*),** the offices of slaveholder and slave 'could not be founded on principles which could be mutually acknowledged'; the only principles which could be mutually acknowledged would be principles that ensured that all did at least as well or better than in the relevant situation of equal liberty.

2.6.5 However, Rawls had changed his model by the time of *Theory* where, as explained in Chapter 1, he assumed a state of general egoism as the relevant situation of equal liberty (§§1.9.32 – 1.9.40). With the state of general egoism as the 1.9.32 position it is by no means clear that both slave and slaveholder would not be prepared to mutually acknowledge principles which permitted the existence of their respective offices. As I suggested in Chapter 1 (§§1.10.1 – 1.10.6), it might be rational for someone who knew that they certainly occupy the office of slave if expediency required such offices to exist, to still prefer a classical utilitarian Leviathan over the alternative of a state of general egoism. So Justice as Fairness might allow the classical principle of utility to be mutually acknowledged by the slave and slaveholder, given the revised assumptions of *Theory*.

2.6.6 If it turned out to be the case that the parties in the revised original position of *Theory* would choose the principle of utility, then there might turn out to be no divergence between the recommendations of Justice as Benevolence and Justice as Reciprocity. This point can be put with reference to Rawls's claims in **Passage 2t (*J as F 2*).** The argument of that passage can be analysed as:

1. If justice were derivative from a higher order executive efficiency it could not account for the fact that slavery is always unjust.

2. Slavery is always unjust

Therefore

3. Justice is not derivative from a higher order of efficiency.

2.6.7 But premise 2 depends on another claim that Rawls wrote a few lines prior to **Passage 2t (*J as F 2*)** that '[a]s that office [i.e. the office of slaveholder] is not in accordance with principles that could be mutually acknowledged, the gains accruing to the slaveholder, supposing them to exist, cannot be accounted as in *any* way mitigating the injustice of slavery.'[1] Premise 2's claim that slavery is unjust, then, depends on Rawls's argument that the office of slaveholder and slave would not be in accordance with principles that could be mutually acknowledged. As I maintained above (§2.6.5), the assertion that the offices in question would not be mutually acknowledged is questionable, given the baseline of general egoism that Rawls is committed to in *Theory.*

2.6.8 If the classical principle of utility turned out to be the choice of the parties in the original position of *Theory*, then Premise 2 of the argument would be false according to the lights of Justice as Fairness, and the conclusion, 3, would not follow. This would open up the possibility that justice might *both* be derivative from a higher order of efficiency *and* derivative from Justice as Fairness. Then, if Justice as Fairness turned out to be the appropriate device for constructing principles appropriate to Justice as Reciprocity (as I shall argue it isn't in the next chapter), Justice as Benevolence and Justice as Reciprocity might converge on the same policy recommendations.

2.6.9 The possibility that Justice as Reciprocity and Justice as Benevolence might converge on the same policy recommendations threatens the ability of Justice as Reciprocity to underwrite the three tenets of deontological liberalism. This can be easily demonstrated. In Chapter 1 I explained how Rawls conceived that Justice as Benevolence could lead to the classical principle of utility. And I showed how classical act-

[1]Rawls 1958 p.188. The sentence from 'Justice as Fairness' (2) effectively makes that same assertion as is equivalent to **Passage 2q (*J as F 1*)**'s Sentence [3].

utilitarianism was unable to support any of the three tenets of deontological liberalism, because of its commitment to unlimited aggregation. So if Justice as Reciprocity also led to classical act-utilitarianism, it would be also be unable to underwrite the three tenets of deontological liberalism.

2.6.10 I should here elaborate on how the tenets of deontological liberalism should be understood for the purpose of this thesis, and that is as pertaining to real people, in the real world, at the present time. Our 'common sense convictions concerning the priority of justice' is that *we* 'have an inviolability founded on justice…which even the welfare of every one else cannot override,'[1] not that people should have an inviolability founded on justice in an ideal world. That, at any rate, is the intuition regarding the core tenet of justice that I uphold in Chapter 4.[2]

2.6.11 The great advantage of the first model in this regard is that it would have fulfilled the promise Justice as Reciprocity offered of supporting the three tenets of deontological liberalism conceived in this way. As described above, the model assumes people who are already engaged in practices (§§2.2.3 – 2.2.4) judging 'whether any of them has a legitimate complaint against their established institutions'. Suppose, then, that Rufus T Firefly had established serfdom five years ago. The serfs in Distribution 4 of **Fig 2 (v)** could legitimately complain that Firefly's decision had deprived them of their liberty-right and claim-right to go away at weekends, as serfdom did not meet the mutual advantage condition of Justice as Reciprocity, and they would only have agreed to principles that would meet the mutual advantage condition. They have these rights as Justice as Fairness supposes that people have the rights to whatever would be accorded to them by principles they would choose in the original position.

2.6.11 The explanation of how the first model would uphold the third tenet of deontological liberalism is slightly more complicated. In the Rawlsian scheme of things,

[1] Rawls *T of J Rev* 1999 (pp. 24 – 27)
[2] For this reason, I have imagined the citizens of Freedonia to be real contemporaries of ours.

different views on how best to be altruistic or benevolent constitute different conceptions of the good. Utilitarianism is regarded by Rawls as a reasonable comprehensive doctrine[1]; its conception of the good could be the statement of Peter Singer's quoted in Chapter 1 which held 'that an act is right only if it does at least as much to increase happiness and reduce misery, for all those affected by it, as any possible alternative act.'[2]

2.6.12 Now, the first model was, I believe, neutral between conceptions of the good. This assertion might seem surprising in light of Section 5's examination of whether egalitarian values could be used to choose among the various distributions that are Pareto-superior to an equal distribution of the advantages of social cooperation. But it should be noted that the context of that examination was *internal* to the question of how to interpret the second principle of justice. And the second principle of justice was chosen on morally neutral grounds: as explained at length, the two principles of justice were chosen as the only principles that could meet the condition of mutual advantage.

2.6.13 This point can be illustrated with reference to **Fig 2 (i).** Suppose that it just so happens that the non-agricultural workers are utilitarians and the agricultural workers are prioritarians. Rufus T Firefly happens to be a utilitarian himself in his personal life, but knows that his responsibility is to implement Justice as Reciprocity and that Justice as Reciprocity leads to the two principles of justice. Furthermore, he takes his responsibilities seriously. He might be tempted to select Distribution 2. Such a selection would further his personal aims in two ways: it would promote the greatest aggregate sum of advantages and it would provide more advantages to people who were themselves inclined to promote the greatest aggregate sum of advantages. But he couldn't in conscience do that. His responsibilities as the person in charge of implementing Justice as Reciprocity oblige him to choose whichever distribution best fulfils the requirement of the second principle of justice, whether that turns out to be distribution 1 or 2.

---

[1] See Rawls 1996 p.145

[2] Rawls describes the relationship between comprehensive moral doctrines and conceptions of the good in *Justice as Fairness: A Restatement.* p. 19. His view is that the elements of a conception of the good are 'normally' set within a comprehensive doctrine but presumably do not have to be. A comprehensive moral doctrine that could provide the setting for the prioritarian conception of the good might be Brian Barry's theory of Justice as Impartiality, as developed across *Theories of Justice* and *Justice as Impartiality*.

2.6.14 The last question to be addressed is whether the classical principle of utility could meet the requirements of Justice as Reciprocity under the revised assumptions of *Theory*. It would obviously meet the constraint requirement; any of the conceptions available to the parties in the original position would have to constrain their behaviour in accordance with the governing principles of society. A slightly trickier question is whether it would meet the mutual advantage condition. This condition, as I defined it in Chapter 1, depends on what the relevant situation of equal liberty is.

2.6.15 The third condition that the classical principle of utility would have to meet is the fairness condition. It is beyond the remit of this thesis to come to a firm conclusion over the question of whether the classical principle of utility could meet the fairness condition, though I shall suggest in Chapter 4, that it is at least not obviously unfair.

2.6.16 In conclusion: the argument of this chapter has been that Rawls's original idea that Justice as Benevolence and Justice as Reciprocity would inevitably prescribe different principles of justice in practice, depended on the first model of his theory's assumption that only the two principles of justice would meet the mutual advantage condition and enable citizens to do as well, or better, than they would in the relevant situation of equal liberty. The shift in Rawls's theory to the adoption of a benchmark of general egoism as the relevant situation of equal liberty, by which to measure mutual advantage, opened up the potential routes from Justice as Reciprocity to the utilitarian conceptions of justice.

Chapter 3. One main ground for the two principles of justice – they're not the principle of utility

In this chapter I examine Rawls's attempts to find an alternative argument, rooted in the conception of Justice as Reciprocity, to block the route from Justice as Reciprocity to the utilitarian conceptions of justice, once that route had been opened by Rawls's revision of the assumptions underlying the first model of his essays 'Justice as Fairness' (1) & (2) to the assumptions of the third model of *Theory.*

This chapter explores Rawls's attempts to do that and argues that they were ultimately unsuccessful.

1 Justice as Fairness

3.1.1 Rawls's most famous attempt to come up with an alternative argument, which was to revise Justice as Fairness, I shall swiftly put aside to concentrate on his other efforts. Ideally it should receive more consideration than I can give it here, but it has received much attention elsewhere and was pretty much abandoned by Rawls himself. The major revision of Rawls's theory of Justice of Fairness was to put considerably more weight on the use of the maximin principle to guide the choices of the parties in the original position than he did in the first model. The maximin principle, as Rawls describes, 'tells us to rank alternatives by their worst possible outcomes: we are to adopt the alternative the worst outcome of which is superior to the worst outcomes of the others'[1] Rawls's argument that it would be rational for the parties in the original position to adopt this principle relied largely on two assumptions: firstly, that they were highly risk averse and secondly; that they cared 'little, if anything, for what [they] might gain above the minimum stipend that [they] can, in fact, be sure of by following the maximin rule.'[2] The first of these assumptions would lead the

[1] Rawls *T of J Rev* 1999 p. 133
[2] Rawls *T of J Rev* 1999 p. 134. Rawls gives a third reason for the deployment of the maximin rule which is that the parties in the original position have no way of estimating their probability of

parties in the original position to favour the two principles over the principle of utility for fear they might end up being losers and the second would lead them to the same ranking of preferences because they care little for the extra advantages the principle of utility would bring if they turned out to be winners.

3.1.2 Rawls himself appeared to have accepted the criticism of reliance on the maximin principle, writing in *Justice as Fairness: A Restatement* that

> In contrast to what the exposition in *Theory* may suggest… the reasons for the difference principle do not rest (as K. J. Arrow and J. C. Harsanyi and others have not unreasonably thought) on a great aversion to uncertainty viewed as a psychological attitude (§§34-39). That would be a very weak argument. Rather, the appropriate reasons rest on such ideas as publicity and reciprocity.[1]

3.1.3 It is those ideas of reciprocity that Rawls refers to that are the preoccupation of this Chapter; some of which are essentially the same in *Theory* as in *Justice as Fairness: A Restatement.*

3.1.4 But before leaving Justice as Fairness behind I shall add one criticism of my own that builds on the criticism that I tentatively offered in Chapter 2.[2] The hypothetical contract device of Justice as Fairness was devised with the purpose of constructing principles suitable for the conception of Justice as Reciprocity. The first model's theory of Justice as Fairness had the parties in the original position choosing the two principles of justice because they were the only principles that would guarantee that all the cooperating parties would fare at least as well, or better, than under the first principle alone. And on the assumptions of the first model they would choose those principles because they knew how principles would be likely to affect them, as they knew what position in society they occupied. However, it can be questioned whether that knowledge of the parties would necessarily lead them to choose principles that guaranteed that they fare at least as well or

---

occupying any particular position in society. This has received extensive criticism from commentators who understand decision theory better than I do, and I defer to their judgement.

[1] Rawls 2001 p. xvii

[2] See §2.2.10

better than under the first principle alone. If there were other principles that were *highly likely* to make everyone, including the worst off members of society, than they would be under the first principle alone, but carried a very small risk that they would make the losers worse off than they would be under the first principle, mightn't it be rational for the worst off people to choose those principles instead? If it was, then the Justice as Fairness of the first model would undermine the purpose of Justice as Reciprocity by potentially foisting principles that didn't treat people as free and equal on real people who hadn't really chosen those principles but whom Justice as Reciprocity was supposed to treat as free and equal. Wouldn't it be more in the spirit of Justice as Reciprocity to insist on the first principle of justice directly?[1]

3.1.5 A much stronger criticism along these lines could be directed against the version of Justice as Fairness Rawls puts forward in *Theory.* The veil of ignorance, that was missing from the first model but that plays a major role in *Theory*, prevents the parties in the original position from knowing how principles would be likely to affect them and the people they represent. So the parties in *Theory's* original position would be able to choose principles that were highly likely to have devastating effects on the real people they represent. It is even open to them to choose principles that would make them worse off than they would be in the no-agreement state of general egoism.

3.1.6 Rawls closes off that possibility in *Theory* by insisting that all the conceptions of justice, including the utilitarian ones, guarantee better prospects than a state of general egoism.[2] This is implausible. Robert Nozick has, rightly in my opinion, pointed out that 'the most pessimistically described Hobbesian state of nature' would not be as bad as 'the most pessimistically described future state' if we include 'future ones'.[3] But even if Rawls were right, the fact that it would be *open* to the parties to inflict principles on real people that would potentially make them worse off than they would be in the relevant situation of equal liberty, *if* such states were possible, points to something wrong with Justice as Fairness. It is not designed to ensure that the principles it selects meet the mutual

[1] As Rawls temporarily did in model 1.5 to be examined in Chapter 4 of this thesis.
[2] As explained in Chapter 1 (§§ 1.9.32 – 1.9.40)
[3] See Nozick 1974 p.5

advantage condition. So it is not fit for the purpose of constructing principles suitable for the conception of Justice as Reciprocity.

3.1.7 That nearly concludes my consideration of Justice as Fairness; though I will go on to consider how Rawls's arguments against utilitarianism look as though they don't need the support of Justice as Fairness below (§3.2.3). I now turn to the more important charges against utilitarianism that are the focus of the remainder of this Chapter. These are all located in Chapter 29 of *Theory* which is titled 'Some main grounds for the two principles of justice'. Given the content of Chapter 29 it might equally well have been titled 'One main ground for the two principles of justice – they're not the principle of utility' since the chapter consists almost entirely of criticisms of the principle of utility.

2 The key passage

3.2.1 The context of Chapter 29 of *Theory* has Rawls weighing up the relative appeal to the parties in the original position of his two principles vis à vis the principle of average utility. The claims under scrutiny all relate to the question of the principle of utility's suitability, or otherwise, to the conception of Justice as Reciprocity and are embedded in the long **Passage 3a (*T of J Rev)*** repeated below. This passage is, I believe, essential to understanding the failure of Rawls's ambition to refute the viability of utilitarianism from the perspective of Justice as Reciprocity. My method in arguing against the charges will be a repeat of the method I deployed in Chapter 1; by comparing the claims in the passage of *Theory* side by side with their historical antecedents from Rawls's earlier essays I aim to expose their weakness.

Passage 3a (*T of J Rev)*

> [1] A conception of justice is stable when the public recognition of its realization by the social system tends to bring about the corresponding sense of justice. [2] Now whether this happens depends, of course, on the laws of moral psychology and the availability of human motives. [3] I shall discuss these matters later on (§§75-76). [4] *At the moment we may observe that the*

> *principle of utility seems to require a greater identification with the interests of others than the two principles of justice.* [5] Thus the latter will be a more stable conception to the extent that this identification is difficult to achieve. [6] When the two principles are satisfied, each person's liberties are secured and there is a sense defined by the difference principle in which everyone is *benefited by social cooperation.* [7] Therefore we can explain the acceptance of the social system and the principles it satisfies by the psychological law that persons tend to *love, cherish, and support whatever affirms their own good.*
>
> [8] When the principle of utility is satisfied, however, *there is no such assurance that everyone benefits.* [9] *Allegiance to the social system may demand that some, particularly the less favored, should forgo advantages for the sake of the greater good of the whole.* [10] Thus the scheme will not be stable unless *those who make sacrifices strongly identify with interests broader than their own.* [11] But this is not easy to bring about. [12] The sacrifices in question are not those asked in times of social emergency when all or some must pitch in for the common good. [13] The principles of justice apply to the basic structure of the social system and to the determination of life prospects. [14] *What the principle of utility asks is precisely a sacrifice of these prospects.* [15] Even when we are less fortunate, we are to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of our life. [16] This is surely an extreme demand. [17] In fact, when society is conceived of as a system of cooperation designed to advance the good of its members, it seems quite incredible that some citizens should be expected, on the basis of political principles, to accept still lower prospects for the sake of others. [18] It is evident then why *utilitarians should stress the role of sympathy in moral learning and the central place of benevolence among the moral virtues.* [19] *Their conception of justice is threatened with instability unless sympathy and benevolence can be widely cultivated.* [20] Looking at the question from the standpoint of the original position, the parties would reject the principle of utility and adopt the more realistic idea of designing the social order on a *principle of reciprocal advantage.* [21] We need not suppose, of course, that in everyday life persons never make substantial sacrifices for one another, since moved by affection and ties of sentiment they often do. [22] But such actions are not demanded as a matter of justice by the basic structure of society.[1] [my italics and numbering of the sentences]

3.2.2 Rawls was unable to sustain any of the italicized claims, and I shall attempt to prove this by showing that they were originally predicated on assumptions that Rawls had 'officially' abandoned by the time of writing *Theory.* His abandonment of the assumptions underlying the italicized claims did not, however, prevent him from putting them forward in the original or revised editions of *Theory,* and of continuing to insist that the difference

[1] Rawls *T of J Rev* 1999 pp.154-155

principle was a reciprocity principle, while the principle of utility was not, in *Justice as Fairness: A Restatement*. In fairness to Rawls, there is one definition he gives (amongst three he considers) of 'reciprocity principle', the conditions of which he could justly claim that the principle of utility could not meet. That is 'the standard of reciprocity'[1] which excludes 'that any one worse off than another should be asked to accept less so that the more advantaged can have more'. But while it is true that the principle of utility is prepared to ask those worse off than others to have less so that the more advantaged can have more, the standard of reciprocity is not, I shall maintain, a viable principle of reciprocity.

3.2.3 But before going into the question of whether the individual italicised claims can be sustained it is worth taking a broader view of the passage in the context of Rawls's theory as a whole. Doing so will help support the short argument I gave above (§§3.1.1 – 3.1.6) that Justice as Fairness is not fit for the purpose of constructing principles suitable for Justice as Reciprocity. For many of the claims in the passage would seem to count against the principle of utility directly from the perspective of Justice as Reciprocity. Take for example, the claim of sentence [8] that '[w]hen the principle of utility is satisfied, however, *there is no such assurance that everyone benefits.'* In the context of Chapter 29, the parties are deliberating over which principles they would choose to govern society. Justice as Fairness says that whatever principles they choose are principles fit for the conception of society as a cooperative venture for mutual advantage; fit for Justice as Reciprocity. There is something very fishy then about allowing principles that don't provide an assurance that everyone benefits onto the menu in the first place. Why couldn't the menu just include items that did provide an assurance that everyone benefits? A similar line of argument applies to the claim of sentence [20] that the parties would 'adopt the more realistic idea of designing the social order on a *principle of reciprocal advantage.'* A natural initial response to that line is to think, 'Well of course the parties should adopt a principle of reciprocal advantage – that's what they're there for.' This could easily be followed by a second, slightly more complex, thought, 'Why do they have to be there at all? Isn't whatever makes the principle of reciprocal advantage a principle of reciprocal advantage

[1] Examined below §3.1.32 - §3.1.33

enough to make it a principle of reciprocal advantage without the need for the parties to choose it in order to make it a principle of reciprocal advantage?' The considerations that the principle of utility doesn't provide an assurance that everyone benefits from social cooperation and isn't a principle of reciprocity (whatever that involves) seems to be enough to disqualify it from the perspective of Justice as Reciprocity anyway. To pile rejection by the parties in the original position on top of the poor principle of utility as well seems a bit like kicking the stuffing out of a corpse.

3.2.4 The inclusion of those two claims, then, reinforces the doubts I have been concerned to raise about the need for Justice as Fairness in Rawls's scheme of things. Now, with Justice as Fairness set to one side, we can consider what force the charges Rawls presses in **Passage 3a (*T of J Rev*)**[1] against the principle of utility may have independently of the role they play in Justice as Fairness. If they turn out to be viable, they would appear to pack considerable punch. From the vantage point of Justice as Reciprocity the claim that the principle of utility isn't a principle of reciprocity seems at least highly damaging. And the claim that the principle of utility can provide no assurance that everyone benefits from social cooperation looks as if it has landed a knock-out blow.

3.2.5 Other charges in **Passage 3a (*T of J Rev)*** against utilitarianism appeal from a broader perspective than that of Justice as Reciprocity. The implication of sentences [10], [12], [14], [21] and [22] is that utilitarianism would demand, or impose, *sacrifices* of some people for the sake of others, in contrast to the two principles of justice which would somehow avoid demanding or imposing sacrifices. Take, for example, sentence [14] '[*w*]*hat the principle of utility asks is precisely a sacrifice of these prospects*?' In the context of the passage, which is comparing the relative merits of the principle of utility and the two principles of justice, sentence [14] should be read as implying that the two principles of justice *don't* ask for a sacrifice of life prospects. The charge that the principle of utility imposes avoidable sacrifices on people is likely to make people recoil from such a principle, whether or not they have bought into the idea of Justice as Reciprocity, or Rawls's theory of Justice as Fairness.

[1] pp. 125 – 126.

3.2.6 So the charges pressed in the passage under consideration would seem to count strongly against the principle of utility, independently of the part they play in Rawls's argument that they would lead the parties in the original position to reject the principle of utility.[1] But whatever force they have depends entirely on whether they are ultimately sustainable. Rawls did not, I shall argue, have the means to sustain them but persisted to press them as though in the hope that he would one day come up with such means.

3 There is 'no assurance that everyone benefits' from the principle of utility

3.3.1 It is worth setting Rawls's claim of sentence [8], that '[w]hen the principle of utility is satisfied…*there is no such assurance that everyone benefits'* against the backdrop of a plausible picture of what would likely happen in the real world were a society to adopt the principle of average utility. **Passage 3a (*T of J Rev)*** [2] evokes a picture of social exclusion for some groups, with some members of society doing badly enough to be considered as excluded from the advantages of social cooperation altogether. It is true that the principle of average utility *could, in principle*, lead to some social groups having very low levels of welfare. It could, to take the extreme possibility that Rawls repeatedly emphasises, even justify slavery. But in my opinion it would be highly unlikely to do so, particularly in a society that constrained the principle of average utility by the requirement that a society was well-ordered, so its citizens consciously accepted the standard of the principle of utility. By comparison to the United Kingdom in 2016, to take an example, I would expect the worst off groups to be considerably better off than they are now. Many of the institutional arrangements that perpetuate inequalities, such as an education system that allows those with more money to buy advantages for their offspring, might go. There might be greater investment in education and training, particularly for the least advantaged groups, leading to greater equality of opportunity and consequently greater equality of outcome. The tax system would likely be more redistributive. There may be some

[1] In her excellent and instructive essay 'Rawls and Utilitarianism', which played an important role in the development of the ideas in this thesis, Holly Smith Goldman addresses what she referred to as Rawls's 'extra-contractarian' arguments (Goldman 1980 p.346). These include the claim that utilitarianism sacrifices some people for others.

[2] pp. 125 – 126.

inequalities due to the effectiveness of providing incentives, and these might result in the worst off doing less well than they would under the difference principle, but I would not expect the difference that would make to be huge.

3.3.2 Let us suppose, for the sake of argument, that the principle of utility would have the benign effects that I suggest it might, with the worst off group in the UK being better off than they are now, but not as well off as they would be under the difference principle, and that the UK suddenly endorsed the principle of utility. We can also suppose that the worst off group is the unemployed, and that they are better off than they are now. Higher levels of welfare support are affordable because the unemployed are more motivated to find work in a more equal society. If someone were to claim that the unemployed did not benefit from their cooperation with the system because they weren't as well off as they might be, most people, I think, would meet that claim with incredulity. Even many of those who believe that the situation of the worst-off is unjust because the better off shouldn't need incentives might well be inclined to agree that they still benefit from social cooperation, but not, perhaps, enough. The claim that social cooperation does not benefit them *at all* seems intuitively implausible.

3.3.3 But, despite the intuitive plausibility of that claim, the historical explanation for Rawls's assertion that the principle of utility can provide no assurance depends on his requiring that a society meet the 'sense defined by the difference principle in which everyone is *benefited by social cooperation'* referred to in sentence [6] in order to benefit its citizens.[1] I shall go into this presently but first I shall consider whether defining the notion of benefitting through social cooperation with reference to a benchmark of general egoism would be enough to sustain the claim of sentence [7] that the principle of utility could provide no assurance that everyone benefits through social cooperation.

3.3.4 It can't. At least it can't in Rawls's hands. For as **Passage 1s (*T of J Rev*)[2]** cited in Chapter 1 showed, Rawls was committed, in *Theory,* to the position that by 'choosing one

[1] By 'historical explanation' here, I mean an explanation in terms of the development of Rawls's theory through the essays leading up to Theory.
[2] p. 67

of the other conceptions [including the utilitarian conceptions] the persons in the original position can do much better for themselves than under general egoism.

3.3.5 Lest **Passage 1s (*T of J Rev)*** be regarded as just a one off slip on Rawls's part, I can point out that he subsequently affirmed this position more clearly in his essay 'The Basic Structure as Subject'. In that essay Rawls was more explicit in taking the no-agreement point of general egoism to be his equivalent of the state of nature of traditional social contract theory.

Passage 3b (*BS as S)*

> we can, if we like, in setting up the argument from the original position, introduce the state of nature in relation to the so-called non-agreement point. This point can be defined as general egoism and its consequences, and this can serve as the state of nature. But these conditions do not identify a definite state. All that is known in the original position is that each of the conceptions of justice available to the parties have consequences superior to general egoism.[1]

3.3.6 With that potential definition of benefit out of the way I now turn to considering whether the sentence [6]'s 'sense defined by the difference principle in which everyone is *benefited by social cooperation'* can fare any better. This possibility subdivides into two further possibilities since Rawls offers two senses in which the difference principle may provide a sense in which inequalities are to the benefit of all. Both need to be considered but both fail in the task.

3.3.7 The two senses are distinguished from each other in the following passage

Passage 3c (*T of J Rev)*: Two senses of mutual benefit

> Next, we may consider a certain complication regarding the meaning of the difference principle. It has been taken for granted that if the principle of justice is satisfied, everyone is benefited. One obvious sense in which this is so is that *each man's position is improved with respect to the initial situation*

[1] Rawls 1996 p. 279

> *of equality*…There may be, however, a *further sense* in which everyone is advantaged when the difference principle is satisfied, at least if we make certain assumptions. [my italics][1]

3.3.8 I shall handle these two senses in the opposite order to which they appear in **Passage 3c (*T of J Rev)***. The assumptions behind the 'further sense' in question are the controversial assumptions of 'close-knitness' and 'chain connection', and it is impossible to appreciate how Rawls may have come to regard them as providing a sense in which social cooperation might be to everyone's advantage without delving into the development of Rawls's theory in the essays between his two 'Justice as Fairness' essays and *Theory*. Going through this history will also explain why Rawls was so determined to come up with 'a further sense' in addition to the 'obvious' one.

3.3.9 Chapter 2 argued that in the first model of Justice as Fairness, the only conception of justice that could meet the mutual advantage condition was Rawls's two principles of justice. The 'relevant situation of equal liberty' by which mutual advantage was to be measured was given by the requirements of the first principle of justice. Furthermore, I suggested, Rawls was committed to the simple assumption of the first model which would have made the second principle of justice easily acceptable to all. All would prefer the one unequal distribution that was Pareto-preferred to the equal distribution, and there would be no dispute between which unequal distribution would be preferred to which other as there would be only one.

3.3.10 The problem of defining a standard of comparison by which to measure mutual advantage became more acute when Rawls abandoned the assumptions of the first model. In Chapter 1, I promised to return to examine the implications of **Passage 1n *(DJ 1967)***[2] from 'Distributive Justice' in more detail, and here is the place where I do so. In Chapter 1, I suggested that there were two problems with substituting 'the liberty of the original position' for the equal liberty of the first principle of justice. The first was that it might allow institutions such as slavery to qualify as 'just'. The second was that it would allow utilitarianism to qualify as just according to the lights of Justice as Fairness.

---

[1] Rawls *T of J Rev* 1999 p. 69
[2] p. 62

3.3.11 The 'further sense' that the difference principle provides which may define when any individual may be advantaged by social cooperation should, I shall argue, be understood as evolving largely as a response to the first two of these problems. To show this I need to first set out the context of **Passage 1n *(DJ 1967)*** from 'Distributive Justice.'(1967)

Rawls's first argument for the difference principle

3.3.12 The paragraph immediately preceding **Passage 1n (*DJ 1967*)** concludes with Rawls, having already defined his second principle of justice in the same words as in 'Justice as Fairness', observing that 'it is not clear what is meant by saying that inequalities must be to the advantage of every representative man, and hence our first question.'[1] **Passage 1n (*DJ 1967*)** follows in the next paragraph which begins

Passage 3d (*DJ 1967)*

> One possibility is to say that everyone is made better off in comparison with some historically relevant benchmark. An interpretation of this kind is suggested by Hume.[2]

3.3.13 So the context of Rawls's reference to Hume's benchmark of a state of nature in **Passage 1n (*DJ 1967)*** was, then, that Rawls was considering possible interpretations of his second principle of justice's stipulation that 'inequalities must be to the advantage of every representative man.'

3.3.14 Rawls then goes on to point out that adopting the benchmark of either a 'hypothetical or historical' state of nature might allow a system of slavery to qualify as being to the advantage of the slaves which leads him to dismiss such a benchmark as

[1] Rawls DJ 1967 p.134
[2] Rawls DJ 1967 p.134

'unsatisfactory. For even if all men including slaves are made better off by a system of slavery than they would be in the state of nature, it is not true that slavery makes everyone (even a slave) better off, at least not in a sense which makes the arrangement just.[1]

3.3.15 Because the benchmark of a state of nature does not provide a sense of 'being made better off' by an arrangement that would make that arrangement seem just, Rawls dismisses the 'historical or hypothetical' benchmark of a state of nature as 'simply irrelevant to the question of justice.'

3.3.16 'In fact', he continues, 'any past state of society other than a recent one seems irrelevant offhand, and this suggests that we should look for an interpretation independent of historical comparisons altogether. Our problem is to identify the correct hypothetical comparison defined by currently feasible alternatives.'[2]

3.3.17 So far, then, Rawls has considered the option of a hypothetical and historical state of nature as a possible interpretation of his second principle of justice's stipulation that 'inequalities must be to the advantage of every representative man' to dismiss it as simply irrelevant because such a definition might allow slavery to qualify as just. And it is this that has led to the 'problem' of finding an alternative interpretation 'in terms of currently feasible changes.'

3.3.18 The second alternative Rawls considers is applying 'the well-known criterion of Pareto' to institutions.[3] This criterion he assumes would be chosen in the original position, but does not go far enough as an interpretation of the second principle of justice's 'benefit to all' stipulation because of its indeterminacy. As discussed in Chapter 2 there is a multiplicity of Pareto optimal distributions, all of which would be preferred by some representative men to others, and none of which would be preferred unanimously to all the

[1] This argument appears to me to represent a determination on Rawls's part to maintain the equation of 'justice' with 'advantage' that we saw was a key feature of the first model (§§2.1.1 - 2.1.7) despite the fact that he no longer subscribed to the assumptions that made the equation work in the first model.
[2] Rawls DJ 1967 p. 135
[3] Rawls DJ 1967 p. 135

others.[1]

3.3.19 The third interpretation of the second principle of justice's stipulation that inequalities must be to the benefit of every representative man Rawls considers is 'the difference principle', referred to as such for the first time,[2] which he introduces thus

Passage 3e *(DJ 1967)*: the first description of the difference principle

> There is, however, a third interpretation which is immediately suggested by the previous remarks, and this is to choose some social position by reference to which the pattern of expectations as a whole should is to be judged, and then to maximize with respect to the expectations of this representative man…Now, the one obvious candidate is the representative of those who are least favored by the system of institutional inequalities. Thus we arrive at the following idea…We interpret the second principle of justice to hold that these differences are just if and only if the greater expectations of the more advantaged, when playing a part in the working of the whole social system, improve the expectations of the least advantaged. The basic structure is just throughout when the advantages of the more fortunate promote the well-being of the least fortunate, that is, *when a decrease in their advantages would make the least fortunate even worse off than they are*.[3]

3.3.20 It is not entirely clear what Rawls is referring to as his 'previous remarks'; it could be his reasons for his dismissal of the Pareto-criterion as insufficient or the benchmark of the state of nature as inadequate or a combination of the two.[4] It does not matter much for my argument, as I am confident that whatever Rawls intended exactly by those words, it was a combination of the reasons for rejecting the first two potential interpretations of the benefit stipulation that motivated Rawls's acceptance of the third: the difference principle.

3.3.21 The difference principle solves the indeterminacy problem of the Pareto criterion by

---

[1] See Rawls DJ 1967 pp. 135-137

[2] Rawls (DJ 1967 p. 138) writes that he refers to 'the first part' of his 'second principle' as 'the difference principle.'

[3] Rawls DJ 1967 p.138 my italics and numbering

[4] We can infer this because in the paragraph immediately preceding Passage 1n (p. 63) Rawls wrote, 'But it is not clear what is meant by saying that inequalities must be to the advantage of every representative man, and hence our first question.' (Rawls DJ 1967 p.134)

picking from the multiplicity of Pareto optimal distributions the one that maximizes the expectations of the least advantaged representative man.

3.3.22 But it also looks like it might offer a solution to the problem of the first interpretation posed by Rawls's supposition that Hume's interpretation of the benchmark of the state of nature might allow slavery to be justified. In fact, there are two ways in which the difference principle seems to provide Rawls with a solution to this problem. The first is relatively uncomplicated. If the prospects of the least fortunate must be as great as they could be, then that would appear to rule out the possibility of slavery. A slave class would presumably be required to endure prospects lower than the prospects of the least fortunate would be, if the requirement that the prospects of the least fortunate were maximized was to be met. So this would get around the problem posed by taking the definition of the benchmark of a state of nature by resorting to another definition that hikes the position of the worst off to as high a level as possible – a level that would preclude the possibility of slavery.

3.3.23 If Rawls had left his argument for the difference principle there, he would have addressed two of his problems: the indeterminacy of the Pareto criterion, and the problem that the benchmark of the state of nature might allow slavery to be justified. But he wouldn't quite have succeeded in doing what he set out to do, which was to find a way to address those problems that also stuck to the wording of his second principle of justice's stipulation that 'inequalities must be to the advantage of every representative man.' This may explain why he then felt the need to 'verify that this interpretation of the second principle of justice [i.e. the difference principle] gives a natural sense in which everyone may be said to be made better off.' This second sense is much more complicated than the first but also more important to my argument as historically it leads to the claims in **Passage 3a *(T of J Rev)***.[1] This way depends on two conditions obtaining in the basic structure of society, chain-connection and close-knitness, on top of the difference principle's stipulation that the expectations of the least advantaged are maximized being met. And the combined effect of these three conditions holding is that

[1] pp. 125 – 126.

Passage 3f (*DJ 1967)*

> everyone does benefit from an inequality which satisfies the difference principle, and the second principle as we have formulated it reads correctly. For the representative man who is better off in any pair-wise comparison gains by being allowed to have his advantage, and the man who is worse off benefits from the contribution which all inequalities make to each position below.[1]

3.3.24 This calls for some explanation. Chain connection obtains when an inequality that raises the expectations of a more advantaged social position, *that has the effect of raising the expectations of the least advantaged*, also raises the expectations of all the positions in between. This goes some way towards making the implementation of the difference principle give a natural sense in which everyone may be said to be made better off by an inequality. But it does not go all the way. For suppose the prospects of the worst off could be improved by offering an advantage to a middle position. The requisite pair wise comparison would hold for the occupants of that middle position and all the positions below. Those in the positions below could each be told that if the positions above them did not have an advantage over them they would be worse off, so the advantage of the position above is to the advantage of the position below. But suppose it would be possible to increase the expectations at the top end of the scale without affecting the positions of those below? This would not violate the difference principle's stipulation that the worst off are as well off as possible. Chain connection might still hold, as it might still be the case that increases in the expectations of more advantaged positions in the middle *that raised the expectations of the lowest position* also raised the expectations of those in between. But it would not be the case that 'the man who is worse off' in any 'pair-wise comparison' has gained from inequalities allowed by the difference principle, since those who are worse off than those to whom the increased advantages at the top end of the scale have been allowed are no better off as a result. So the condition of close-knitness is also needed. The second condition of close-knitness holds if 'it is impossible to raise (or lower) the expectation of any representative man without raising (or lowering) the expectations of every other

[1] Rawls DJ 1967 p. 139

representative man.'[1] This condition does not, it should be noted, exclude the possibility of increases in the expectations of more advantaged men which have the effect of *lowering* the position of the worst off class. But such increases would violate the stipulation of the difference principle. So it would be impossible to raise the advantages of a more advantaged position (without violating the difference principle) without also *raising* the expectations of every representative man. The net effect of these three conditions is to ensure that the requisite relationship holds and that the worst off representative man in any pairwise comparison could console himself that the advantage of the position above him was, indeed, to his advantage in that if the more advantaged representative man had any less, the representative man would have less still.

3.3.25 Grasping the mechanics of these two conditions is not particularly important here, but it is important to note two things. Firstly, that this interpretation would seem to provide Rawls with a second way in which the difference principle would solve the problem posed by the alternative of defining the advantageousness of social cooperation by comparison with the benchmark of a state of nature. The only advantages that would be allowed to any better off party would be those that were to the advantage of the positions of those below. But it seems extremely unlikely that the advantages of any positions that benefited from a system of slavery would be to the advantage of the slaves, the worst off group, in this way. So taking this 'sense in which everybody may be said to be made better off' rather than the alternative of using the benchmark of a state of nature might seem to offer Rawls another way out of the problem posed by the fact that a system of slavery might be to the slave's advantage, according to the benchmark of a state of nature interpretation.

3.3.26 The second point to note is that Rawls concedes that '[o]f course, chain-connection and close-knitness may not obtain', in which case '[t]he stricter interpretation of the difference principle should be followed, and all inequalities should be arranged for the advantage of the most unfortunate even if some inequalities are not to the advantage of those in middle positions.' He then goes on to remark 'Should these positions fail, then, the

[1] Rawls DJ 1967 p.139

second principle would have to be stated in another way.'[1] The question this raises is: why was it so important to Rawls to stick to the original wording of his second principle?

3.3.27 In answer to this question it is worth adding my thoughts to comments made by Robert Paul Wolff regarding 'a rather odd characteristic of Rawls's exposition.' Wolff points out that when Rawls realizes that the two principles, as formulated in his first model wouldn't work out as planned, his 'obvious move is to give up his formula, and search instead for a different set of principles that meet the theoretical demands he wishes to place on them.' [2] Instead of doing this, Rawls looks for some way to make his new principle, the difference principle, fit the original wording of his second principle of justice. I would suggest that the explanation for this oddity lies largely in the original formula's advantages from the perspective of Justice as Reciprocity. The two advantages of the first model were as follows. Firstly, that it conflated the distinct ideas of 'inequalities being to everyone's advantage, and 'social cooperation being to everyone's advantage'. And secondly, that it made 'social cooperation (under the rules of a particular institution) being to everyone's advantage' a necessary and sufficient condition for the institution in question's being just. This conflation becomes more explicit in the passages from 'Distributive Justice: Some Addenda' that I shall consider shortly. And it lies, as my tracing of the history of the claims of **Passage 3**a (***T of J Rev)*** will eventually show, behind the assertion of line [5] that the difference principle provides a sense 'in which everyone is benefited by social cooperation' and the implicit claim of line [7] that the principle of utility can provide 'no such assurance that everyone benefits.'

3.3.28 Although Rawls presented his discussion of the three possible interpretations of the second principle of justice's stipulation that inequalities are to the benefit of all as simply an attempt to ascertain its meaning, I have described it, and examined it, as his first argument for the difference principle. I hope that my examination of it in the paragraphs above (§3.1.12 - §3.1.26) has justified this treatment. Rawls first arrived at the difference

[1] Rawls DJ 1967 pp. 139 -140
[2] Wolff 1977 pp. 57-58

principle, I have argued, in an attempt to come up with a new principle of mutual benefit that would, as his principles of justice aimed to do in the first model, allow us to reconcile our intuitions as to what institutions were just, with those that the principle defined as beneficial.

3.3.29 It should be noted that, read this way, this argument for the difference principle is entirely independent of the arguments for the difference principle from Justice as Fairness as they would later be presented in *Theory*. These are firstly, the argument that it would be appropriate for the parties in the original position to use the maximin rule, and secondly, the argument that underlies **Passage 3a** (***T of J Rev)***, that the parties would reject the principle of utility in favour of the two principles of justice, due to the latter's greater propensity for ensuring a stable society. So a similar question arises in relation to this argument as arose in Chapter 2. If the principles of justice can be motivated directly from considerations of Justice as Reciprocity, then why is there a need for Justice as Fairness?

'Distributive Justice: Some Addenda.' (1968)

3.3.30 'Distributive Justice: Some Addenda', as the name suggests, builds on the analysis of the principles of justice that Rawls put in 'Distributive Justice'. The reason for my analysing it here is that it is in 'Distributive Justice: Some Addenda' that Rawls first predicates some of the claims of *Theory's* **Passage 3a (*T of J Rev)*[1]** on the 'further sense' in which the difference principle meets the second principle of justice's stipulation that 'inequalities must be to the advantage of every representative man.'

3.3.31 These claims include: a) the claim of line [6] that the difference principle satisfies 'the psychological law that persons tend to love, cherish, and support whatever affirms their own good'; b) the claim of line [5] that the difference principle provides 'a sense in which everyone may be said to be made better off' and the related claim of line [7] that the principle of utility can 'provide no such assurance that everyone benefits.' It is also in

[1] pp. 125 - 126

'Distributive Justice: Some Addenda' that Rawls starts to refer to this sense as defining a condition of mutual benefit or reciprocity, which he continues to do in *Theory*. And he does so in a passage, the first part of which, is repeated almost word for word in *Theory*. I have separated the passage from 'Distributive Justice: Some Addenda' into two, in order to clearly distinguish between the two different criterions of reciprocity it contains,

Passage 3g (*DJ:SA)*: the standard of reciprocity

> A further consideration in support of the difference principle is that it satisfies a reasonable standard of reciprocity. Indeed, it constitutes a principle of mutual benefit, for, when it is met, each representative man can accept the basic structure as designed to advance his interests. The social order can be justified to everyone and in particular to those who are least favoured. [4] By contrast with the principle of utility, *it is excluded that any one worse off than another should be asked to accept less so that the more advantaged can have more.* This condition seems an essential part of the notion of reciprocity and the difference principle fulfils it where utilitarianism does not.[1]

Continuing with

Passage 3h (*DJ:SA)*: the condition of mutual benefit

> It is necessary, however, to consider how the condition of mutual benefit is satisfied. Consider any two representative men A and B, and let B be the one who is worse off. Actually, since we are most interested in the comparison with the least favoured man, let's assume that B is this individual. Now clearly B can accept A's being better off since A's advantages have been gained in ways that improve B's prospects. If A were not permitted to win his better position, B would be even worse off than he is.[2] [my italics]

3.3.32 One of Rawls's tendencies as a writer, pronounced enough to be described as a trait, is to present seemingly similar, but significantly different, claims as though they were one and the same. The passages just quoted provide a good example of this. The 'condition of mutual benefit' described in **Passage 3h (*DJ:SA)*** depends on the conditions of chain

[1] Rawls DJ:SA 1968 p.169
[2] Rawls DJ:SA 1968 p.169

connection and close-knitness holding, while the 'standard of reciprocity' described by the italicised sentence in **Passage 3g (*DJ:SA)*** doesn't, and I shall refer to them by those titles from now on in order to distinguish them. If **the condition of mutual benefit** holds then so will **the standard of reciprocity**, but the converse is not true. Recall the example of §3.1.25 where chain connection held but close-knitness didn't and it was possible to raise the prospects of better off groups without affecting the prospects of the worse off ones, and impossible to raise the prospects of the worse off groups. Then the condition of mutual benefit of **Passage 3h (*DJ:SA)*** would not be fulfilled while the standard of reciprocity of **Passage 3h (*DJ:SA)*** might be. A possible explanation for this is that Rawls was confident at this stage that these conditions would hold.[1]

3.3.33 The standard of reciprocity and the condition of mutual benefit differ only marginally from the similar conditions put forward in 'Distributive Justice' in the italicized passage at the end of **Passage 3e *(DJ 1967)*[2]**, and the second sense that the difference principle gives 'in which everyone may be said to be made better off.' But in 'Distributive Justice' Rawls was seemingly just concerned to provide a correct interpretation of the second principle of justice's stipulation that inequalities are to the advantage of every representative man. Here he has upped his game against utilitarianism with his remark that the standard of reciprocity is 'an essential part of the notion of reciprocity and the difference principle fulfils it where utilitarianism does not.'

---

[1] Someone might be tempted at this point to read Rawls as defining the principle of mutual benefit as holding just when the advantages of more favoured individuals are to the advantage of the *least* advantaged individual rather than between any two representative men, in which case the assumptions of chain connection and close-knitness need not hold. Two factors count against this reading. Firstly, he clearly states that the condition of mutual benefit should apply between *any* two representative men, he is just taking the least representative man as an example;,and secondly, his reconstruction of **Passage 3g** in the original edition of *Theory* includes a caveat about chain connection. There he wrote 'A further point is that the difference principle expresses a conception of reciprocity. It is a principle of mutual benefit. We have seen that, *at least when chain connection holds,* each representative man can accept the basic structure as designed to advance his interests.' (Rawls 1971 p. 102 my italics). This passage was substantially altered in the revised edition of *Theory,* so it no longer contained either the caveat or the comparison between representative men A and B. This revision may have been prompted by extensive criticism of the relevant passage in Robert Nozick's *Anarchy, State and Utopia.* On a related issue, concerned with whether Rawls should be read as accepting the principle that 'inequality is unjust, unless it benefits the worst-off group,' see Derek Parfit's *Equality or Priority* page 119 and footnote 59.

[2] p. 135

3.3.34 Rawls now goes on to make a series of claims about the relative suitability or unsuitability of the difference principle and the principle of utility to society conceived of as a cooperative venture for mutual advantage that are direct corollaries of the ones in **Passage 3a (*T of J Rev)***, writing

Passage 3i *(DJ:SA 1968*) 'affirming our good'

> [1] The difference principle should be acceptable, then, both to the more advantaged and to the less advantaged man. [2] The *principle of mutual benefit* applies to each increment of gain for the more favoured individual, a unit increase, so to speak, that improves the situation of this individual being allowed provided that it contributes to the prospects of the least fortunate.[1] [3] The *principle of reciprocity* applies each step of the way, the increments for the better situated continuing until the mutual benefit ceases. [4] It is evident that, in general, the principle of utility does not satisfy the *principle of reciprocity*; there is no *definite sense in which everyone necessarily benefits from the inequalities* that are authorized by the utilitarian conception. [5] It seems irrelevant to say that everyone is better off than he would be in a state of nature, or if social cooperation were to break down altogether, or even that all are better off in comparison with some historical benchmark. [6] We want to be able to say that as the social system now works, the inequalities it allows contribute to the welfare of each.
>
> [7] Now the fact that the two principles of justice embody *this reciprocity principle* is important for the stability of this conception. [8] A conception of justice is stable if, given the laws of human psychology and moral learning, the institutions which satisfy it tend to generate their own support, at least when this fact is publicly recognized. [9] Stability means that just arrangements bring about in those taking part in them the corresponding sense of justice, that is, a desire to apply and act on the appropriate principles of justice. [10] Assuming as a basic psychological principle that *we tend to cherish what affirms our good* and to reject what does us harm, all those living in a basic structure satisfying the two principles of justice will have an attachment to their institutions regardless of their position. [11] This is the case since *all representative men benefit from the scheme.* [12] In a utilitarian society, however, this is not guaranteed; and therefore to the extent that this psychological principle holds the principle of utility is likely to be a less stable conception.[2]

[1] The same point, I think, applies here as in the footnote above.
[2] Rawls DJ:SA 1968 p. 170-171 my italics and numbering of the sentences.

3.3.35 It is impossible, I think, to say whether Rawls was meaning to invoke the standard of reciprocity or the condition of mutual benefit with his references to 'the principle of reciprocity', 'the principle of mutual benefit' and 'this reciprocity principle' here, and I doubt very much that Rawls would have had a clear answer himself. What is clear is that Rawls, in this passage, intends one or other of them to sustain the claims: a) of line [4] that 'there is no definite sense in which everyone necessarily benefits from the inequalities that are authorized by the utilitarian conception; b) of line [10] that a basic structure ordered by the principles of justice, because they embody the reciprocity principle (whatever that may be) will lead those living under it to have an attachment to their institutions due to the psychological principle that we tend to cherish what affirms our own good, and c) of line [11] that all representative men in a basic structure which is ordered by the principles of justice will, because the principles embody the reciprocity principle, benefit from the scheme. What is also readily apparent is the implied claim d) of lines [12] and [4] that representative men in a utilitarian society may not benefit from the principle of utility, just because it doesn't embody this reciprocity principle.

3.3.36 **Passage 3a (*T of J Rev)*[1],** whose claims form the primary focus of the investigation of this chapter, is obviously in large part, a re-edit of **Passage 3i *(DJ:SA)*. 3a**'s line [7] is the analogue of **3i**'s line [10]. **3a**'s line [8] is **3i**'s line [4] and **3a**'s line [6]'s roots lie in the implicit claim of **Passage 3i *(DJ:SA)*** that the standard of reciprocity or condition of mutual benefit provide a sense in which the difference principle ensures that everyone benefits from social cooperation. This will receive more explanation very shortly.

3.3.37 What appears more clearly in **Passage 3i (*DJ:SA)***, than anywhere else, including 'Distributive Justice', is that Rawls, in this second stage of his theory, was concerned that his stipulation that inequalities be to everyone's advantage should provide a viable alternative to defining mutual advantage in comparison to the benchmark of a state of nature, which he again dismisses as irrelevant.

3.3.38 What is also apparent from this passage is just how powerful a weapon, in the

[1] pp. 125 – 126.

difference principle's arsenal, Rawls, at that time, regarded the condition of mutual benefit to be. All the advantages ascribed to the two principles of justice in the second paragraph of **Passage 3i (*DJ:SA*)**[1] and all the disadvantages there ascribed to the utilitarian conception of justice are attributed to the former's meeting this condition ('this reciprocity principle') and the latter's failing to. So, only if a conception of justice meets this condition should it be regarded as 'affirming one's good', leading people to cherish it and have an attachment to their institutions regardless of their position. And only if this condition is met can all representative men be regarded as benefiting from the scheme of cooperation. Furthermore, the implication is that any conception of justice that falls short of meeting this condition should actually be construed as harming them.

3.3.39 Before proceeding to the question of whether Rawls continued to sustain the claim that the difference principle would meet the condition of mutual benefit in *Theory*, we should consider whether, even if it did, the condition of mutual benefit would be as powerful as Rawls portrays it. And I think it should be fairly obvious from the observations just above that he has it punching well above its weight. Not only would the worst off position have to be as well off as they could be, but no further advantages could be allowed to any other group that didn't redound to the advantage of the group below for all the groups to even be considered as benefiting from the cooperative scheme. Only if these conditions were met would everyone's good be affirmed. And, as just mentioned above, the further implication is that Rawls might then have even regarded groups as being harmed by conceptions of justice which didn't meet this condition. Counterexamples to these claims should not be hard to find, and indeed they're not.

3.3.40 First, suppose that the worst off cooperating group in society were impossible to help beyond a certain point, so it would be possible to raise the prospects of all groups bar the worst off, after the position of the worst off had been maximised. This would break the condition of mutual benefit since none of the advantages of the better off groups would count as being to the advantage of the worst off in the requisite way. According to Rawls, the worst off group should then not be considered to have benefited from the cooperative

[1] p. 143

scheme, their good would not have been affirmed and they might even be regarded as harmed by such a move. But that is absurd. They are no worse off than they would have been had the prospects of the more advantaged groups not been raised.

3.3.41 Second, consider the example I introduced in §3.3.1 of the United Kingdom in 2016 converting to utilitarianism, where the worst off group in society are a lot better off than they are now but could be still better off under Rawls' difference principle. The same argument applies. It just seems intuitively wrong to regard the people who would be worst off after this conversion, who would be considerably better off than the worst off group in the United Kingdom are now, to have not benefited from social cooperation.

3.3.42 We can conclude from the above that, although fulfilment of the condition of mutual benefit might give the difference principle some advantages in terms of ensuring stability, it is inadequate as an objective requirement that society conceived of as a cooperative venture for mutual advantage must meet in order to be considered such, which is how Rawls presents it here.

A Theory of Justice

3.3.43 At any rate, by the time of *A Theory of Justice* Rawls had reverted to the position he took in 'Distributive Justice', where he conceded that should the conditions of close-knitness and chain–connection fail to hold, meaning that the difference principle would no longer meet the condition of mutual benefit, the expectations of the worst-off group should still be maximised.[1] In fact, he went further than before by introducing the idea of 'leximin', allowing that once the expectations of the worst off group had been maximised the position of the next worst off group could be maximised (although they didn't improve the position of the group below) and so on up the scale until the position of the best off group could be maximised so long as it wasn't at the detriment of anyone lower down the scale. But he still continued to deploy the condition of mutual benefit as a possible extra string to the difference principle's bow just in case the conditions of close-knitness and

[1] Rawls *T of J Rev* 1999 p.70

chain-connection did hold, referring to it variously as 'the principle of mutual advantage'[1] ,'a further sense in which everyone is advantaged when the difference principle is satisfied'[2], 'a sense in which everyone benefits when the difference principle is satisfied'[3], 'a conception of reciprocity', 'a principle of mutual benefit'[4], 'the condition of mutual benefit'[5], 'the criterion of mutual benefit'[6] and, perhaps, in **Passage 3a (*T of J Rev*)**[7] as 'a sense defined by the difference principle in which everyone is benefited by social cooperation.'[8]

3.3.44 As noted above (§3.3.26), in 'Distributive Justice' Rawls suggested that '[s]hould these positions fail, then, the second principle would have to be stated in another way.' And, true to his word, he did state it in another way that the difference principle could be assured of fulfilling with or without those conditions obtaining, though it is fair to speculate that he can't have been too satisfied with it. This is the sense whereby the difference principle ensures that 'each man's situation is improved with respect to the initial arrangement of equality'.[9]

3.3.45 I'm now in the position to clarify further the remarks I made earlier (§3.3.35) about it being impossible to say which 'sense defined by the difference principle in which everyone is benefited by social cooperation' Rawls intended to lie behind the assertions of **Passage 3a (*T of J Rev*).** The surrounding claims about the difference principle's affirming one's good, where utilitarianism, by implication, wouldn't, and the principle of utility providing no assurance that everyone benefits from the scheme of cooperation clearly have their historical roots in **Passage 3i *(DJ:SA 1968*)** from 'Distributive Justice: Some Addenda' where Rawls presented the condition of mutual or standard of reciprocity as the *only* contender for the sense in question. So that sense must be read as lying behind the

[1] Rawls *T of J Rev* 1999 p. 69.
[2] Rawls *T of J Rev* 1999 p. 69
[3] Rawls *T of J Rev* 1999 p. 70
[4] Both Rawls *T of J Rev* 1999 p. 88
[5] Rawls *T of J Rev* 1999 p. 103
[6] Rawls *T of J Rev* 1999 p. 89
[7] p. 143
[8] Rawls *T of J Rev* 1999 pp. 154-155
[9] Rawls *T of J Rev* 1999 p. 69

claims of the **Passage 3i *(DJ:SA 1968)*[1]**. But how to tell which is supposed to lie behind the same claims in *Theory* when that sense has been downgraded to a secondary one, a new one has been asserted as the main one, and the author has given no indication as to which one he has in mind? The question is unanswerable.

3.3.46 But we can ask the following two questions: firstly, whether the primary sense would provide a better definition of social cooperation's being to one's advantage than the secondary one that we have already seen to be inadequate; and, secondly, whether this primary sense would provide better support for the surrounding claims than the secondary one.

3.3.47 Rawls hypothesizes that the parties in the original position would choose the difference principle in two stages. First, they would choose to distribute all income and wealth equally. Then, they would choose to allow inequalities that worked to everyone's advantage, and would select the distribution that maximised the position of the worst off out of those unequal distributions. The parties don't know which generation they belong to. So the initial arrangement of equality referred to as the baseline from which the difference principle ensures that everyone is advantaged is whatever distribution would have resulted from extending a principle of flat equality through the past, at whatever time the choice from the original position is imagined to take place and the veil of ignorance is lifted.

3.3.48 This does not constitute a good definition of social cooperation's being to one's advantage, and its inadequacy as a definition of that notion can be illustrated, I think, by seeing it in the light of the evolution of the second principle of justice's stipulation that inequalities be to everyone's advantage that I have been concerned to outline in this Chapter, and in particular the implicit version that he let slip, as though in error, in **Passage 1k *(DJ 1967)*[2].** That version, as I interpreted it, lay halfway between the version of his first model which compared practices where everyone enjoyed equal freedom with practices that allowed pareto-preferred advantages measured in economic terms, and his final version

[1] p. 143
[2] p. 53

where the benefits of economic inequalities are measured in comparison to a benchmark of economic equality. The hybrid version had the parties in the original position swapping their imagined *equal freedom* for the economic advantages that social cooperation would bring, and was tantamount, I suggested, to imagining advantages to be defined against the benchmark of a hypothetical state of nature defined as the no-agreement point.[1] That, I suggested, might have provided a good definition of the notion of social cooperation's being to one's advantage. The difference principle as presented in *Theory* has to be seen as a reluctant compromise on Rawls's original ambitions for his second principle of justice.

3.3.49 But although the primary sense in which the difference principle ensures that inequalities are to the advantage of all is inadequate as a definition of social cooperation's being to one's advantage, we can ask, parenthetically, whether the principle of utility would, as Rawls claims, provide no assurance that everyone benefits if measured by that standard. And the answer to that depends on what economic theory one holds. According to some of the assumptions Rawls makes that underlie his theory, it is not at all clear that he should regard the principle of utility as failing according to this standard. For he holds that the main reason that inequalities might be required for the position of the worst off in society to be maximized is that the more advantaged need incentives in order to be more productive. It is quite possible that the deadening effect of allowing no incentives on the economy would mean that flat equality would make the position of the worst off group worse than they would be under either the difference principle or the principle of utility.[2] In fact, although he does not explicitly address this question, there is an indication that Rawls himself would take this view revealed in the diagrams he uses to illustrate the difference between the principle of utility and the difference principle. These take the origin, 0, to represent the hypothetical state in which all goods are distributed equally, and

[1] Although I have read Rawls in *Theory* as interpreting his second principle of justice's stipulation that inequalities be to the benefit of all as given by the primary sense in which the difference principle ensures that inequalities are to everyone's advantage, there are still echoes of the broader sense in which he had earlier conceived of it surviving in *Theory.* For example, he remarks 'that there are indefinitely many ways in which all may be advantaged when the initial arrangement of equality is taken as a benchmark', (p.65) before going on to consider systems of natural liberty, liberal equality and democratic equality as possible interpretations of 'everyone's advantage'. But the difference principle forms part of the conception of democratic equality. The other conceptions are not nearly so 'egalitarian' or supportive of the worst off.

[2] I am indebted to Mike Otsuka for this point.

describe a curve where the principle of utility gives the worst off representative man less than the difference principle would but more than equality would.[1] So even if the primary sense in which the difference principle ensures that inequalities are to the advantage of all is taken to lie behind Rawls' claims in **Passage 3a (*T of J Rev)*[2]**,** it appears that he could not sustain the claim that the difference principle defines a sense in which social cooperation is to everyone's advantage while the principle of utility provides no such assurance.

3.3.50 As showed at the start of this chapter, the most coherent definition Rawls deploys of social cooperation's being to everyone's advantage is by comparison to a state of nature taken to be one of general egoism. And he is implicitly committed to taking the principle of utility to ensure that social cooperation *is* to everyone's advantage by this comparison. So we can conclude that Rawls cannot sustain the claim in question.

4 Sacrifices for others

3.4.1 The purpose of this section is not just to undermine Rawls's argument from Justice as Fairness – that is, that the parties in the original position would not favour the principle of utility because of the sacrifices it requires – but also to mitigate the widely held view that there must be something wrong with utilitarianism just because it requires sacrifices. This is held by philosophers who are not at all sympathetic to Rawls's theory of Justice as Fairness. To take a case in point, Robert Paul Wolff, who, as we have already seen, is a trenchant critic of Justice as Fairness, wrote in *Understanding Rawls*, that, 'Utilitarianism, in even its most sophisticated and complicated versions, countenances the sacrifices of some persons to the happiness of others.'[3]

3.4.2 The term 'sacrifice' is an emotionally charged one, and I think many people reading

[1] Rawls *T of J Rev* 1999 pp.66-67. He repeats what is essentially the same diagram in *Justice as Fairness: A Restatement* 2001 p.62 suggesting his views had not changed, though he does not there specify what the origin represents.
[2] Rawls *T of J Rev* 1999 pp.154-155
[3] Wolff 1977 p 12

that quotation of Wolff's would instantly recoil against the idea of endorsing a moral theory that would 'sacrifice' some people for the 'happiness of others' and be inclined to go looking for an alternative moral theory with less 'abhorrent implications'[1] Several of the claims of **Passage 3a (*T of J Rev*)**[2] lodge an appeal to that intuitive response, and so are likely to have had an influence in prejudicing people against utilitarianism, that is entirely independent of the role they play in Justice as Fairness.

3.4.3 My argument of this section will not succeed in demonstrating that utilitarianism does not countenance the sacrifice of some persons for the happiness of others. But it will succeed in showing that Rawls's two principles of justice would also sacrifice some persons for the sake of others. This is something that Rawls does not explicitly deny, but its denial is implicit in his pressing of the claims against utilitarianism. I shall also suggest that justice *requires* sacrifices of persons for the sake of others, and those sacrifices are sacrifices of natural liberty, though it is beyond the remit of this thesis to prove this conclusively.

3.4.4 What I will be able to prove is that Rawls was unable to establish any basis for a claim that his principles do not sacrifice some persons for others. I may as well signal in advance that my argument does not, I think, show Rawls in a particularly flattering light. For by tracing the historical development of the 'sacrifice' claims of **Passage 3a (*T of J Rev)***, I shall show that he certainly considered what I have suggested is the correct position regarding justice and sacrifices of liberty, but he was not prepared to acknowledge it, at least not in *Theory*.

3.4.5 Before embarking on the task of tracing the history of Rawls's sacrifice claims I should first verify that **Passage 3a (*T of J Rev)*** does indeed put forward the claims I am attributing to Rawls. This will require my showing firstly, that Rawls is making a claim that the principle of utility requires sacrifices for others *per se*, rather than sacrifices of the less advantaged for the sake of greater benefits to the more advantaged. And secondly, that

[1] Wolff 1977 p 11
[2] pp. 125 - 126

he also made the implicit claim that his two principles would avoid sacrifices *per se*.

3.4.6 The first of these is necessary because some commentators have interpreted Rawls as just charging the principle of utility with requiring sacrifices of the less advantaged for the sake of greater benefits to the more advantaged. Following David Brink, I shall call these 'bottom-up' sacrifices.[1] There are good reasons for such an interpretation. Some of the sentences of **Passage 3a *(T of J Rev)*[2]** are worded in such a way that they could be plausibly interpreted as an objection to 'bottom-up' sacrifices; for example, the claim of sentence [15] that '[e]ven when we are less fortunate, we are to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of our life.' And, pertinently, my comparison in Chapter 1 of **Passage 1s *(T of J Rev)*[3]** from *Theory* with its progenitor **Passage 1j (*DJ 1967)***, interpreted **Passage 1s *(T of J Rev)*** as containing an objection to bottom up sacrifices. Furthermore, the standard of reciprocity is an objection to bottom up sacrifices.

3.4.7 However, other sentences in **Passage 3a *(T of J Rev)*** cannot be plausibly interpreted as objections to bottom up sacrifices, and I shall set these out in the order that they appear in **Passage 3a *(T of J Rev).***

3.4.8 First there is sentence [9]: 'Allegiance to the social system may demand that some, particularly the less favored, should forgo advantages for the sake of the greater good of the whole.' Next there is sentence [10] 'Thus the scheme will not be stable unless those who make sacrifices strongly identify with interests broader than their own'. Thirdly, there is the couplet of sentences [13] and [14], 'The principles of justice apply to the basic structure of the social system and to the determination of life prospects. What the principle of utility asks is precisely a sacrifice of these prospects.' Finally, there is the last pair of sentences of **Passage 3a *(T of J Rev)***, [21] and [22], 'We need not suppose, of course, that in everyday life persons never make substantial sacrifices for one another, since moved by affection and ties of sentiment they often do. But such actions are not demanded as a matter of justice by

[1] Brink 1993 p.6
[2] pp. 125 – 126.
[3] p. 67

the basic structure of society.'

3.4.9 It is difficult to see how Rawls could sustain these claims, without also conceding that similar charges could be levelled against his two principles of justice. Rawls does not specify how sacrifice is to be construed, but one possibility would be to measure it as the difference between what someone would have in the well-ordered society governed by one conception of justice compared with what they would have in the well-ordered society governed by an alternative conception of justice. However, this measure would construe both the two principles of justice and the principle of average utility as requiring sacrifices. Those who would fare better under the principle of average utility would, in the terms of [13] and [14], undergo a 'sacrifice' of 'life prospects' in comparison to what they would have under the principle of average utility.[1]

3.4.10 But the tone of the sentences just cited, coupled with the context of the passage imply that Rawls did not concede that the two principles of justice would be subject to similar charges. The context is that the parties are choosing between alternative conceptions of justice, and the charges in question are presented as reasons for them favouring the principles of justice over the principle of average utility. These charges should not provide such reasons if the two principles of justice were open to the same objections.

3.4.11 Rather than consider any alternative possible measures of sacrifice, I turn now to the history of Rawls use of the term 'sacrifice' in the essays preceding Theory. I'm confident that the light of that history will reveal Rawls's claims that the principle of utility would require sacrifices *per se*, that his principles could somehow avoid, to be unsustainable.

3.4.12 The first reference to sacrifice was in the 'Justice as Fairness' (1) in the context of the parties in the first original position choosing the second principle of justice. Rawls asserted that 'Each person will, however, insist on a common advantage, for none is willing

[1] It should be borne in mind that by this stage in *Theory*, Rawls had narrowed the choice of utilitarian conceptions of justice down to the principle of average utility, making it a 'fixed sum game', so one person's loss under the principle of average utility would equate to another person's gain under the two principles of justice and vice versa.

to sacrifice anything for the others.'[1]

3.4.13 From the perspective of the parties in the first original position, choosing the principle of utility could have been read as requiring a willingness to sacrifice 'something' for others. The context of Rawls's assertion in the paragraph above assumes that the parties in the first original position have already accepted a baseline of equal liberty because, 'Since there is no way for anyone to win special advantage for himself, each might consider it reasonable to acknowledge equality as an initial principle.' So as measured from the first model's baseline of equal economic distribution and equal liberty (as I explained in Chapter 2, they amounted to the same thing in the first model), the parties in the first original position would not be required to make sacrifices by the two principles of justice but would by the principle of utility.

3.4.14 In 'Justice as Fairness' (2), Rawls repeated the same assertion as in (1) but remarked in addition, 'I do not want, therefore, to be interpreted as assuming a general theory of human motivation: when I suppose that the parties are self-interested, and are not willing to have their (substantial) interests sacrificed to others,'[2]

3.4.15 These remarks are the origin of the theme that the principle of utility required sacrifices for others that the two principles of justice would avoid; a theme that Rawls appears to have been determined to pursue, despite the fact that the underlying assumptions of his model changed so as to make the charge no longer viable.

3.4.16 Another feature of the first model that is relevant to my argument of this section is that whilst Rawls acknowledged that acting in accordance with the principle of fair play involved a 'restriction of liberty', his first model provided him with the means to claim that by another sense of liberty, the two principles of justice avoided a loss of liberty while the principle of average utility did not. This would be because, as explained in Chapter 2, the two principles of justice would provide all with an equal or greater liberty, while the

[1] Rawls 1957 p.657
[2] Rawls 1958 p.175

principle of utility would provide them with a less than equal liberty, 'liberty' then being measured in terms of the economic benefits that a certain assignment of rights would bestow on the participants in a practice or cooperative venture.

3.4.17 However, Rawls was to change his position regarding both liberty and 'sacrifice' when he abandoned the first form of the model. In 'The Sense of Justice' (1963), he was prepared to concede that both Justice as Fairness and the principle of utility would require sacrifices of interests

Passage 3j (*S of J 1963*): sacrifice in accordance with justice as fairness

> justice as fairness is correct in viewing each person as an individual sovereign, as it were, none of whose interests are to be sacrificed for the sake of a greater balance of happiness but rather only in accordance with principles which all could acknowledge in an initial position of equal liberty.[1]

3.4.18 This is a particularly significant passage. It does not present the objection to utilitarian sacrifices as the sentences analysed in paragraph §3.4.8 do, and as the quotation from Wolff does (§3.4.1), as objections to sacrifices *per se*. Instead, it presents an objection to sacrifices for the purpose of promoting the greatest balance of happiness rather than sacrifices that are in accordance with principles which all could acknowledge in an initial position of equal liberty, i.e. than in accordance with Justice as Fairness.

Sacrifices: the only viable position

3.4.19 This, I think, is the only viable position that Rawls considered regarding sacrifices. That position is that Justice as Reciprocity requires sacrifices. What Rawls hoped would distinguish the two principles of justice from the principle of average utility, in **Passage 3j (*S of J 1963*),** in his essay 'The Sense of Justice' was that the two principles of justice

[1] Rawls S of J 1963 p. 304

would be selected by Justice as Fairness while the principle of utility wouldn't. Then the criticism of the principle of utility from the perspective of Justice as Reciprocity would make sense. It would not be that the principle of utility required sacrifices *per se* that was what was wrong with it. It would be that the principle of utility required sacrifices just for the purpose of promoting the greatest happiness whereas the two principles of justice would only allow sacrifices in accordance with principles that people would choose in the original position. The principles of justice would, then, require sacrifices, but sacrifices which were fair.

3.4.20 The explanation for Rawls presenting this objection this way is that his essay 'The Sense of Justice' assumed that the two principles of justice had already won the argument from Justice as Fairness, and defeated the principle of utility. Therefore, Rawls could uphold a basic contrast between two conceptions of justice; the two principles of justice on the one hand, which were favoured by Justice as Fairness, and the principle of utility on the other, which aimed to promote the greatest happiness.

3.4.21 However, the context of **Passage 3a *(T of J Rev)*[1]** is very different. The context of **Passage 3a *(T of J Rev)*** does not assume that the argument from Justice as Fairness has been won. Rather, it is the very question of who will win that contest which is at stake, and the fact that the principle of average utility imposes sacrifices on some – sacrifices that the two principles of justice would, by implication, avoid - is a factor that is supposed to help determine the result of that contest.

3.4.22 This is just one of several places in *Theory* where Rawls did not take the context of the argument from Justice as Fairness properly into account. It is particularly evident in sections from Part Three of *Theory*. I will reveal this fully in the forthcoming sections addressing the claim of sentence [20] from **Passage 3a *(T of J Rev),*** that the parties would reject the principle of utility in favour of a principle of reciprocal advantage, and the claims of [4], [5] and [19] regarding the greater stability of the two principles in comparison with the principle of average utility.

[1] pp. 125 - 126

3.4.23 **Passage 3i**[1] allowed that citizens in a society ordered by the two principles of justice would have to undergo sacrifices of their interests. But it didn't quite equate those sacrifices with a loss of liberty. Rawls came closest to doing this in the essay following 'The Sense of Justice', 'Legal Obligation and the Duty of Fair Play' (1964)

Passage 3k *(LO & DFP 1964)*: principle of fair play

> The principle of fair play may be defined as follows. Suppose there is a mutually beneficial and just scheme of cooperation, and that the advantages it yields can only be obtained if everyone, or nearly everyone, cooperates. Suppose further that cooperation requires a certain sacrifice from each person, or at least a loss of liberty.[2]

3.4.24 What is remarkable about Rawls's position in these two passages is the shift they display from Rawls's position of the first model which held that the two principles of justice would not require any sacrifice of interests or loss of liberty, to a position where he acknowledges that they would require sacrifices or a loss of liberty. Rawls's new position retains one feature of the old model in equating sacrifices of interests with loss of liberty. The concession that the principles of justice would require a sacrifice of liberty (as I shall assume it can reasonably be interpreted) was prompted, I would speculate, by Rawls's realization that the first model's goal of having the two principles of justice continuous with each other, with the second providing all with more liberty than the first, was unachievable. But there were two unfortunate side-effects of Rawls's changing underlying assumptions, from the point of view of Justice as Reciprocity. The first is that they required more weight to be placed on the choice of the parties in the original position. The second is the removal of the very feature that made the principles of justice most attractive from the point of view of Justice as Reciprocity, which was that only they, in contrast to the principle of utility, would ensure that all did as well or better than in a position of equal liberty.

3.4.25 Rawls was to try one more approach to this theme before *Theory,* and that was in

[1] p. 143
[2] Rawls LO& DFP 1964 p.122

'Distributive Justice', with the problematic argument of **Passage 1k (*DJ 1967*)**[1], which entertained the idea that the principle of utility would not provide all with compensating advantages for their loss of the liberty of the original position. This idea can be seen as a brief attempt to recapture the idea of the first model that the principle of utility would disadvantage people with respect to a baseline of equal liberty. It is not hard to understand the appeal of this idea from Rawls's perspective. If it had proved viable then he would have been able to sustain lines of argument along the lines of those of those of **Passage 3a** (***T of J Rev*)** under scrutiny, as I demonstrate with a fictional scenario just below.

3.4.26 Imagine, then, that Freedonia was an island where everyone enjoyed the equal liberty of the original position, and that in Freedonia, everyone was able to advance their interests to some extent, without the need for any rules of social cooperation. But there existed another island, Utilitaria, which already had an indigenous population who were on the brink of extinction as they were unable to look after themselves. The governor of Utilitaria calls for the people of Freedonia to come to migrate on mass to Utilitaria, where they would be placed under the rule of institutions governed by the principle of average utility. If the free people of Freedonia were to accept his call then at least some, if not all, of them would end up worse off than they would be remaining in Freedonia, in part because of the burden of supporting the indigenous population of Utilitaria. We can also suppose that both the principle of average utility and the principle of classical utility would favour the migration to Utilitaria as the natives of Utilitaria would gain more than the free people of Freedonia would lose. If the Freedonians were to answer the governor of Utilitaria's call, then the sacrifice claims of **Passage 3a *(T of J Rev)*** could be upheld.

3.4.27 Take first, sentence [9] of **Passage 3a *(T of J Rev)***: 'Allegiance to the social system may demand that some should, particularly the less favoured, forgo advantages for the sake of the greater good of the whole.' Allegiance to the social system of Utilitaria would require some to forgo the advantages they enjoyed under the equal liberty of Freedonia for the sake of the greater good of the whole, the whole being the combined population of the two islands. Then take sentence [10] 'Thus the scheme will not be stable unless those who

[1] p. 53

make sacrifices strongly identify with interests broader than their own'. If we assume that the people of Freedonia had the option of giving up and returning home and resuming their previous lifestyles, then the scheme would not be stable unless those who lost through their move to Utilitaria identified with the interests of the native Utilitarians. Third, the claim of sentences [13 and [14], 'The principles of justice apply to the basic structure of the social system and to the determination of life prospects. What the principle of utility asks is precisely a sacrifice of these prospects.' The principle of utility, which would direct the Freedonians to set sail for Utilitaria would, in so doing, ask at least some them to sacrifice the greater prospects they would have enjoyed staying put. Finally, [21] and [22], 'We need not suppose, of course, that in everyday life persons never make substantial sacrifices for one another, since moved by affection and ties of sentiment they often do. But such actions are not demanded as a matter of justice by the basic structure of society.' The principle of utility, and the governor of Utilitaria, would have demanded substantial sacrifices on the part of at least some Freedonians, as measured by the loss of their advancement of their interests they would have enjoyed in Freedonia.

3.4.28 It should also be observed that Wolff's criticism that 'Utilitarianism, in even its most sophisticated and complicated versions, countenances the sacrifices of some persons to the happiness of others,' could then be interpreted in such a way as to be upheld. If we take utilitarianism to be 'act utilitarianism' which, as I explained in Chapter 1, requires everyone, no matter what position they may be in, to act so as to maximize utility, then act utilitarianism would demand the sacrifice of the people of Freedonia. This could be interpreted as a sacrifice of their natural liberty, for the sake of promoting the happiness of the natives of Utilitaria, through its demand that they obey the governor of Utilitaria's call for assistance.

3.4.29 I believe that even if Rawls had persisted with the idea that the principle of utility would not provide all with compensating advantages for their loss of the liberty of the original position he would have faced serious difficulties in translating the idea that the principle of utility would not compensate all the parties in the original position for their loss of the 'liberty of the original position' into the idea that the principle of utility would impose sacrifices on the citizens of the well-ordered utilitarian society. However, I do not

need to address this question since, as shown previously, Rawls abandoned this idea in favour of assumptions according to which the principle of utility would not sacrifice people by comparison with what they would have in a state of natural liberty, that being equivalent to a state of general egoism.

3.4.30 The correct position regarding the justice or injustice of sacrifices and their relationship to liberty and interests is, I believe, the one Rawls considered in his essays 'The Sense of Justice' and 'Legal Obligation and the Principle of Fair Play' which held that justice required a sacrifice of interests or of liberty. He would also have been right to clearly equate the two, as he nearly did in 'Legal Obligation and the Principle of Fair Play.' Justice requires a sacrifice of the liberty to pursue one's interests. I cannot offer conclusive proof that this is the right position but I can offer some considerations in its favour.

3.4.31 Imagine the case of a gifted lawyer who turns down the opportunity to work as consigliere for the Mafia out of respect for the law, in a society that is well-ordered by the two principles of justice. She might be said to have forgone advantages or made a sacrifice. She could have made millions with the mafia, but is instead settling for hundreds of thousands and the sacrifice or sum of advantages forgone is measured by the loss of earnings over her life due to this choice.

3.4.32 Her cooperation with the law of might also be said to involve a 'loss of liberty', to use the terms Rawls used in his description of the principle of fair play in **Passage 3k (*LO & DFP*)**.[1] The liberty in question would be the natural liberty to do whatever she wanted in pursuit of her goals. If we assume that she would have quite liked to reap the rewards of working for the mafia were it not for her sense of justice directing her to constrain such self-interested behaviour, then she could be regarded as having lost her liberty in the relevant sense.

3.4.33 Or, to take a more high-minded example, we could imagine a practitioner of a religion that abhors blasphemy against their religion. One of the tenets of their religion

[1] p. 157

demands the destruction of any material that it blasphemous towards their most important prophets. But a rival religion has an evangelical commitment to denigrating the prophets of their rivals, in order to win converts, which extends to publishing material that the practitioners of the first religion would deem blasphemous. There would appear here to be a clash between upholding the basic liberty of conscience and the basic liberty of freedom of speech, both of which are protected by the first principle of justice. It would, I think, be in the general spirit of Rawls's first principle of justice to resolve the dispute on the side of the blasphemous religions. But this resolution would still require restraint on the part of the practitioner of the blasphemy-abhorring religion. If this resolution is right the practitioner of the blasphemy-abhorring religion should refrain from destroying the 'blasphemous' literature of the other religion. Although any restraint on the part of the practitioner of the blasphemy-abhorring religion might be described as done in the name of freedom, that 'freedom' being the freedom guaranteed by the basic liberties, it should also, I think, be interpreted as requiring a sacrifice of the natural liberty to pursue one's interests by whatever means one can, and as a sacrifice of the interests of the practitioner of the blasphemy-abhorring religion. Justice in the society that is well-ordered according to the two principles of justice would still require sacrifices both of liberty and of 'life prospects', to use the term Rawls utilizes in sentences [13] and [14] of **Passage 3a (*T of J Rev*).**

3.4.34 So that, I think, is the only viable position that Justice as Reciprocity can take. Just behaviour requires a sacrifice of *liberty* in return for the benefits provided for a similar sacrifice of liberty undertaken by others. Whether or not the principle of utility is unjust from the perspective of Justice as Reciprocity will depend on other factors than that it imposes sacrifices on its citizens. All the conceptions of justice (apart from the non-starter of general egoism) would.

5 Principle of Reciprocal Advantage

3.5.1 There is only one sentence in **Passage 3a (*T of J Rev*)** where Rawls maintains that the principle of utility is not a principle of reciprocal advantage and that is Sentence [20]. The question of what potential power that charge might have has already been addressed above (§3.2.3). So the question that remains is: *what principle was Rawls referring to by*

'principle of reciprocal advantage'?

3.5.2 There are three possibilities. The first is that it is another term for the 'condition of mutual benefit' or 'the standard of reciprocity'. I argued above that the 'condition of mutual benefit' was an inadequate condition of reciprocity and indeed, that Rawls conceded it to be such (§3.1.43), so the condition of mutual advantage needs no further consideration here. The second is that it is the principle of reciprocity considered in Chapter 2 and in Rawls's essay 'Justice as Reciprocity' (1971). The third is that it is a 'reciprocity principle'[1] that Rawls evokes in Part 3: Section 76 of *Theory,* that may, or may not, prove to be distinct from the standard of reciprocity already discussed. I shall consider the question of whether it is an alternative principle of reciprocity in Section 5 after swiftly disposing of the second possibility just raised; that it is the principle of reciprocity Rawls referred to in his essay 'Justice as Reciprocity' (1971).

3.5.3 In 'Justice as Reciprocity' (1971) Rawls wrote that '[t]he principle of reciprocity requires of a practice that it satisfy those principles which the persons who participate in it could reasonably propose for mutual acceptance under the circumstances and conditions of the hypothetical contract.'[2] All those conceptions should presumably be regarded as 'reasonable to propose'. So to preserve the spirit, if not the letter, of the 'principle of reciprocity' of 'Justice as Reciprocity' it seems fit to translate principle of reciprocity as requiring of institutions[3] that they are in accordance with the principle or principles that the parties would eventually choose from those that could be reasonably proposed. On this interpretation, Rawls had no right to assert in **Passage 3a (*T of J Rev)*** 'the parties would reject the principle of utility and adopt the more realistic idea of designing the social order on a *principle of reciprocal advantage.'* For the context of **Passage 3a (*T of J Rev)*** has the parties choosing which principles would fit the bill of Justice as Fairness. If they chose the principle of utility then it would be the principle of reciprocity.

3.5.4 The sections surrounding 75 and 76 also go into Rawls's claim of **Passage 3a (*TofJ***

[1] Rawls *T of J Rev* 1999 p.437
[2] Rawls 1971 p.208. As already cited in **Passage 2n (*J as R 1971*)** Chapter 2.
[3] 'Institutions' is here substituted for 'Justice as Reciprocity' (1971)'s reference to 'practices'.

Rev) that the well-ordered utilitarian society, in contrast to the two principles of justice, would have to rely on sympathy and benevolence in order to be stable. The claim that the well-ordered utilitarian society would depend on sympathy in contrast to the well-ordered society that instantiated the two principles of justice is one that is reiterated at length in *Justice a Fairness: A Restatement,* providing further reason for examining it in detail.

6 Reciprocity versus sympathy

3.6.1 I repeat the three relevant sentences of **Passage 3a (*TofJ Rev)*** here for convenience.

> [4] *At the moment we may observe that the principle of utility seems to require a greater identification with the interests of others than the two principles of justice.* [18] It is evident then why *utilitarians should stress the role of sympathy in moral learning and the central place of benevolence among the moral virtues.* [19] *Their conception of justice is threatened with instability unless sympathy and benevolence can be widely cultivated.*

3.6.2 These claims look clearly relevant to the main argument of my thesis, that Justice as Reciprocity can lead to the classical principle of utility. But can they be sustained? Would the principle of average utility have to depend on sympathy any more than the two principles of justice would?

3.6.3 Rawls's argument that it does turns out to depend on the aforementioned 'standard of reciprocity', though one has to read Sections 75 and 76 of *Theory* very carefully to work that out.

3.6.4 Here is the passage from Section 76 of *Theory* where Rawls mentions goes into the question at length.

> **Passage 3l (*T of J Rev*): the well-ordered society paired with the principle of utility.**

> [1] We can confirm this suggestion by considering the well-ordered society paired with the principle of utility. [2] In this case, the three psychological laws would have to be altered. [3] For example, the second law now holds that persons tend to develop friendly feelings toward those who with evident intention do their part in cooperative schemes publicly known to maximize the sum of advantages, or the average well-being (whichever variant is used). [4] In either case the resulting psychological law is not as plausible as before. [5] For suppose that certain institutions are adopted on the public understanding that the greater advantages of some counterbalance the lesser losses of others? [6] Why should the acceptance of the principle of utility (in either form) by the more fortunate inspire the less advantaged to have friendly feelings toward them? [7] This response would seem in fact to be rather surprising, especially if those in a better situation have pressed their claims by maintaining that a greater sum (or average) of well-being would result from their satisfaction. [8] No *reciprocity principle* is at work in this case and the appeal to utility may simply arouse suspicion. [9] Thus the attachments generated within a well-ordered society regulated by the utility criterion are likely to vary widely between one sector of society and another. [10] Some groups may acquire little if any desire to act justly (now defined by the utilitarian principle) with a corresponding loss in stability.
>
> [11] To be sure, in any kind of well-ordered society the strength of the sense of justice will not be the same in all social groups. [12] Yet to insure that mutual ties bind the entire society, each and every member of it, one must adopt something like the two principles of justice. [13] It is evident why the utilitarian stresses the capacity for sympathy. [14] Those who do not benefit from the better situation of others must identify with the greater sum (or average) of satisfaction or else they will not desire to follow the utility principle. [15] Now such altruistic inclinations no doubt exist. [16] Yet they are likely to be less strong than those brought about by the three psychological laws formulated as reciprocity principles: and a marked capacity for sympathetic identification seems relatively rare. [17] Therefore these feelings provide less support for the basic structure of society.[1]

3.6.5 The clues that it is 'the standard of reciprocity' doing the bulk of the work here is in Sentences [5] to [8] and [14]. The essence of Rawls's argument is that the worst-off in society wouldn't require sympathy in the well-ordered society paired with the two principles of justice, because then they would know that the advantages allowed to the more advantaged would be helping them (or at least not harming them).

3.6.6 Before considering the strength of this argument, I will set aside a red herring, which is provided by Sentence [2]'s assertion that 'the psychological laws would have to be

[1]Rawls *T of J Rev* 1999 p.437

altered' in the well-ordered society paired with the principle of utility and by [15] and [16]'s implication that 'the three psychological laws formulated as reciprocity principles' couldn't apply to the well-ordered society. These might be taken to imply that Rawls had another viable principle of reciprocity in mind, which is independent of the standard of reciprocity. He didn't, and once more there is a historical explanation for the appearance that he had, which I shall explain after first demonstrating that the second psychological law is entirely consistent with the principle of utility.

3.6.7 The second psychological law in question reads

Passage 3m (*T of J Rev*)

> given that a social arrangement is just and publicly known by all to be just, then this person develops ties of friendly feeling and trust toward others in the association as they with evident intention comply with their duties and obligations, and live up to the ideals of their station.[1]

3.6.8 Now the context of **Passage 3l (*T of J Rev*)** is that Rawls is comparing the well-ordered society paired with the principle of utility *on the assumption that it has been chosen by the parties in the original position* with the well-ordered society pared with the two principles of justice on the assumption. If the principle of utility had been chosen by the parties, the social arrangements under it would, to use the words of **Passage 3m (*T of J Rev*),** be just and known to be just. So there would be no need to alter the wording of the second law as Rawls asserts there is in Sentence [2] of **Passage 3l (*T of J Rev*).**

3.6.9 The historical explanation behind the implication that the psychological laws are incompatible with the principle of utility lies in the fact that they were first formulated in Rawls's essay 'The Sense of Justice' (1963). As I observed above (§§3.2.20 – 3.2.22) this essay assumed that the two principles of justice had been selected by Justice as Fairness, and that the principle of utility hadn't.[2] In the context of 'The Sense of Justice' (1963),

[1] Rawls *T of J Rev* 1999 pp. 429-430

[2] Rawls repeatedly fails to take the fact that it is an open question which principles, the two principles of justice or the principle of utility, the parties would ultimately choose in Part 3 of

then, the implicit claim of Sentence [16], that the three psychological laws formulated as reciprocity principles could only apply to the well-ordered society paired with the two principles of justice, and not to the well-ordered society paired with the principle of utility would have arguably made more sense. The term 'reciprocity' could have carried with it the same connotations as the term carried in the context of 'the principle of reciprocity' of Rawls's essay 'Justice as Reciprocity' (1971) or as used in 'Justice as Fairness' (1)[1], which meant roughly 'as selected by Justice as Fairness'. But it is inappropriate for 'reciprocity' to carry that connotation in the context of the argument of Part 3 of *Theory*, when it is the very question of which principles would be selected by Justice as Fairness which is at stake.

3.6.10 So my conclusion of the last three paragraphs (§§3.6.3 – 3.6.9) is that any appearance that there is an alternative reciprocity principle to the 'standard of reciprocity' doing any work in the argument of **Passage 3l (*T of J Rev*)** should be disregarded**.** Either the psychological laws formulated as reciprocity principles should be read as compatible with the principle of utility or they should be taken to rely covertly on the standard of reciprocity.[2]

Theory. So he wrote, '[t]hus in arguing further for the principles of justice as fairness, I should like to show that this conception is more stable than other alternatives.' (Rawls *T of J Rev* 1999 p.399) He clearly means 'the two principles of justice' by 'the principles of justice as fairness' while his theory of Justice as Fairness is open to selecting one of the other alternatives to live up to the designation. In a similar vein, he writes, '[b]ut a decision in the original position depends on a comparison: other things equal, the preferred conception of justice is the more stable one. Ideally, we should compare the contract view with all its rivals in this respect, but as so often I shall only consider the principle of utility.' (Rawls *T of J Rev* 1999 p.436) Here 'the contract view' is used to refer to the two principles of justice and not to the principle of utility, when, as the first sentence concedes, the decision in the original position is yet to be made.

[1] As examined in Chapter 2 (§§2.4.1 – 2.4.9)

[2] One commentator who appears to have embraced the idea that there is a psychological law of reciprocity which would attach to the two principles of justice but not to the principle of average utility is Samuel Scheffler. Scheffler (2003) writes, 'Rawls argues…that because his principles embody an idea of reciprocity or mutual benefit, and because reciprocity is the fundamental psychological mechanism implicated in the development of moral motivation, the motives that would lead people to internalize and uphold his principles are psychologically continuous with developmentally more primitive mechanisms of moral motivation… By contrast, utilitarianism does not embody an idea of reciprocity. If people are to be stably motivated to uphold utilitarian principles and institutions, even when those principles and institutions have not worked to their advantage, the capacity for sympathetic identification will have to be the operative psychological mechanism.' As it has been a major concern of my thesis to demonstrate, in the light of Rawls's endorsement in *Theory* of the benchmark of a state of general egoism as the benchmark by which to

3.6.11 I return now to the substantive issue of whether there really is a difference in kind between the motive 'sympathy' which the well-ordered society paired with the principle of utility would have to call on in order to be stable, and the motive of 'reciprocity' that the well-ordered society paired with the two principles of justice could depend on. To address this question it will help to focus on the pair of sentences '[13] It is evident why the utilitarian stresses the capacity for sympathy. [14] Those who do not benefit from the better situation of others must identify with the greater sum (or average) of satisfaction or else they will not desire to follow the utility principle' from **Passage 3l (*T of J Rev*).**

3.6.12 To simplify the question let us consider, as Rawls so often does, how the claims of these sentences would apply to the worst-off group in society. And I shall take my consideration of these claims again against the backdrop of the 'plausible picture of what would likely happen in the real word were a society to adopt the principle of average utility' that I set out at the start of this chapter (§§3.1.1 – 3.1.2). So the worst off group in the United Kingdom would be better off than the worst of group is now but not as well off as they would be if the United Kingdom had adopted the two principles of justice. Rawls's claim of sentence [14] then, as applied to that example, might be that the unemployed would only stick to the straight and narrow rather than resorting to a life of crime if they had sympathy for the working taxpayer they might harm by their criminal activity. They must identify with the greater sum (or average) of satisfaction (in this case a lower tax burden for the working taxpayer) or else they will not desire to follow the utility principle (by embarking on a life of crime). There is also an implicit claim contained within these sentences, which is that the unemployed would not need to rely on sympathy under the two principles of justice because they realized that the relative advantages of the working taxpayer redounded to their advantage, and were not won at their expense.

3.6.13 However, a claim analogous to sentence [14] could be pressed against the two principles of justice. Under the two principles of justice, the average working taxpayer, we can assume, would have less than she would have under the principle of average utility.

---

gauge 'mutual advantage', Scheffler's assertion that utilitarian 'principles and institutions have not worked to [the citizens in a well-ordered utilitarian society's] advantage' begs the question.

Her conformity to the principles of justice might require her to not subscribe to illegal tax avoidance schemes. So it could be argued, analogously to Sentence [14], that under the two principles of justice 'those who do not benefit from the difference principle's requirement that the worst-off be as well off as possible (by comparison to what they could have under the principle of average utility) must identify with the plight of the worst off group in society or else they will not desire to follow the two principles of justice.'

3.6.14 The point of raising the analogous claim that might be put on behalf of the working taxpayer who might have more under the principle of utility than under the two principles of justice is to query why sympathy for the worst off would not need to be fostered in the well-ordered society governed by the two principles of justice, just as sympathy for the better off would need to be fostered in the well-ordered society paired with the principle of utility?

3.6.15 Rawls does have an answer to that question, but not, to my mind, a convincing one. In *Justice as Fairness: A Restatement* (2001), Rawls supposes that the more advantaged may be willing to accept losses imposed on them by the two principles of justice in comparison to what they might have under an alternative arrangement of society, because 'they are mindful of a deeper idea of reciprocity implicit in the difference principle'[1] in addition to its fulfilment of the standard of reciprocity.[2] This deeper idea implicit in the difference principle is that 'social institutions are not to take advantage of contingencies of native endowment, or of initial position, or bad luck over the course of life except in ways that benefit everyone, including the least favoured.'[3] So Rawls's argument is that the more advantaged do not need to rely on *sympathy* to motivate them to help the least favoured as they accept that *reciprocity* requires them to use their advantages in ways that benefit everyone. However, a counter-example can be given to question whether acceptance of

[1] Rawls 2001 p.126
[2] Rawls does not refer to 'the standard of reciprocity' in those words in *Justice as Fairness: A Restatement.* Instead he refers to it as 'the following reciprocity condition: those who are better off at any point are not better off to the detriment of those who are worse off at that point.' (Rawls 2001 p.124) This stipulates the same conditions as the standard of reciprocity examined in this chapter (§§3.3.32 – 3.3.33).
[3] Rawls 2001 p.124.

this deeper idea of reciprocity must really favour the difference principle over the principle of average utility. Suppose a working, and conscientious, taxpayer had a choice between voting for three political parties. The Blue party would cut his tax bill. The Brown party would increase his tax bill and spend it on helping the worst off group in society, who happen to be a small minority of congenitally disabled people who are very expensive to help. The Purple party would increase his tax bill by the same amount as the Brown party but invest it in education in deprived areas, and in so doing provide substantial benefits for a large, relatively badly off group, who aren't the worst off group. The principle of utility would favour his voting for the Purple party. If this taxpayer voted for the Purples rather than the Blues, this could be due to an acknowledgement that he had an obligation to use his advantages in ways that are to everyone's advantage despite the fact that he would not be helping the worst off group in society. So this deeper idea of reciprocity does not seem to necessarily favour the two principles of justice over the principle of average utility. If the more contingently advantaged are to be persuaded to care more about helping the worst off group than other groups, it seems to me that a 'tie-breaker' would be needed once 'the deeper idea of reciprocity' that directs one to put one's talents to everyone's advantage has been acknowledged. That tie-breaker will blur any clear distinction between sympathy and reciprocity.

3.6.16 My conclusion to this section is that Rawls failed to establish the claims of **Passage 3a (*T of J Rev)*** that the principle of utility must rely on sympathy rather than reciprocity. I repeat them here: [4] *At the moment we may observe that the principle of utility seems to require a greater identification with the interests of others than the two principles of justice* and [18] It is evident then why *utilitarians should stress the role of sympathy in moral learning and the central place of benevolence among the moral virtues.* [19] *Their conception of justice is threatened with instability unless sympathy and benevolence can be widely cultivated.*

3.6.17 There is some truth in the claim, I think, that it would be especially hard to motivate the less advantaged to comply with rules for the sake of greater benefits to the more advantaged, but this difference in motivation is a matter of degree rather than kind.

3.6.18 Two final remarks can be made regarding the claim that utilitarians must stress the role of sympathy and the central place of benevolence in moral learning. The first is that the principle of utility would be able to accommodate any instability that might be caused by losses incurred by the worst-off in comparison with how they would fare *within the principle itself.* If the worst-off threaten to undermine society by not complying with the rules of society the principle of utility says 'give them more'. Not to do so, and to permit a dangerous and restless underclass to exist, would be not to maximize utility. So the principle of utility would provide benefits to those who could cause sufficient trouble when it would deny them those benefits if they didn't have the potential to cause trouble

3.6.19 The second remark is that where our sympathies lie is a contingent matter. Many people are concerned about issues such as global warning which threaten to destroy the planet (or at least make large parts of it uninhabitable). It is fair to suppose that those people would prioritize such issues over making the worst off as well off as possible. If more people develop sympathies or benevolence in an outward looking direction, then principles of distribution, such as the principle of utility, that take those sympathies into account, may provide more stability than those that don't.

Concluding Remarks

3.7.1 In the opening section of this chapter I described the position that Rawls's theory of Justice as Reciprocity was in due to the failure of his first model. The effect of shifting to an endorsement of a benchmark of general egoism had potentially unblocked the route from Justice as Reciprocity to the principle of utility.

3.7.2 This chapter has looked at Rawls's attempts to erect new blocks and argued that they all failed. In Section 3 I argued that Rawls did not have the means to sustain his claim that not everyone would benefit through cooperation in the well-ordered society paired with the principle of utility. In Section 4 I argued that Rawls's implicit claim that the principle of utility would impose sacrifices on people while the two principles of justice would avoid imposing sacrifices on people was false. Importantly, I also suggested that Justice as Reciprocity would require sacrifices from the citizens of the well-ordered society: sacrifices

of their natural liberty and the extra benefits they might acquire through the unconstrained pursuit of their conception of the good. In Section 5, I addressed the question of what Rawls meant by his claim that the parties would reject the principle of utility in favour of a principle of reciprocity and suggested that the only principle of reciprocity that the principle of utility could be claimed to violate was the 'standard of reciprocity'. In Section 6 I looked at Rawls's claim that the principle of utility must rely on the motive of sympathy, while the two principles of justice could rely on the different motive of reciprocity and argued that he had not succeeded in sustaining that claim.

3.7.3 So, at the end of this chapter, Rawls's theory of Justice as Reciprocity is in pretty much the same condition as it was at the beginning. According to the assumptions of Theory, the utilitarian conceptions of justice would meet the mutual advantage condition of Justice as Reciprocity; and there appears to be no other condition rooted in the conception of Justice as Reciprocity that would rule out utilitarianism.

Chapter 4. Reconstructing Rawls

In this Chapter I attempt to show that the classical principle of utility can be reconciled with the conception of Justice as Reciprocity. The course my argument will take roughly follows the outline given of this chapter in the 'Chapter by Chapter Outline of Thesis'.[1] So **Section 1** argues that, given the assumptions of *Theory* and *Justice as Fairness: A Restatement,* classical act-utilitarianism would be fully reconcilable with Justice as Reciprocity. In **Section 2** I take account of the fact that the assumptions of *Theory* are wrong, and that accommodating that fact would lead to modifying classical act-utilitarianism by exempting those for whom the demand to maximize utility afforded sufficiently dismal life prospects from the obligation to maximize utility. The resulting conception of justice I call 'Reciprocal Classical Utilitarianism'. In **Section 3** I consider the implications of Justice as Reciprocity's reconcilability with 'Reciprocal Classical Utilitarianism' for the three tenets of deontological liberalism. I show that it motivates the core tenet, but not the other two.

The central argument of this thesis will be concluded by the end of **Section 3,** as my goal of establishing that the classical principle of utility is reconcilable with the conception of Justice as Reciprocity will have been achieved. But, in what might initially be viewed as something of a side issue, in **Section 4** I examine Rawls's widely known and highly influential 'separateness of persons' objection to utilitarianism. This has often been interpreted as an objection to utilitarianism that is entirely independent of Rawls's conception of society as a cooperative venture for mutual enterprise with its concomitant conception of Justice as Reciprocity. But, in what I believe is an original interpretation of the objection, I put forward a case that it should be interpreted as essentially the same as 'the exclusion of aggregation' of Rawls's first model that was examined in Chapter 2: that is, it is an objection to violations of the mutual advantage condition.[2] This fits well with my thesis's theme of demonstrating that Rawls's fundamental objection to utilitarianism

[1] This thesis p. 14

[2] First defined in §2.1.18

was that it was incompatible with Justice as Reciprocity; the appearance that the separateness of persons' objection to utilitarianism is a distinct objection is, I propose, illusory.

1 From Justice as Reciprocity to Classical Act-Utilitarianism

4.1.1 At the close of Chapter 1, it was suggested that it would be rational for anyone, even if they knew they would be in the worst off group in society, to sign a contract pledging allegiance to a classical utilitarian Leviathan, even if they also knew the classical utilitarian Leviathan would be prepared to subject them to the most abject of slaveries. It would be rational for them to sign such a contract as their expectations under the classical utilitarian Leviathan would still be better than in the default position of general egoism. In which case, Justice as Reciprocity would have proceeded, via Justice as Fairness, to have instituted classical act utilitarianism. In this section I review that suggestion in light of the findings of Chapters 2 and 3.

4.1.2 The three requirements of Justice as Reciprocity that classical act utilitarianism would have to meet were **the constraint requirement, the mutual advantage condition** and **the fairness condition**. The fact that classical act-utilitarianism would meet the first requirement can be swiftly demonstrated. All conceptions of justice, bar general egoism, would require constraint on the part of the cooperating members of society.

4.1.3 Classical act-utilitarianism would meet the mutual advantage condition on Rawls's assumptions of the third model of *Theory*. This has already been demonstrated in Chapter 1. On Rawls's assumptions all the other conceptions of justice, including the utilitarian ones, offer better expectations than general egoism.[1] And general egoism represents the no-agreement point for Justice as Reciprocity, which I am taking to be the relevant situation of equal liberty by which to gauge whether the mutual advantage condition has been met.

4.1.4 The most difficult condition to assess is the fairness condition. I won't be able to

[1] See **Passage 1s (*T of J Rev)*** p. 67.

show that the classical principle of utility is the conception of justice that is uniquely fair. But I will be able to offer a couple of considerations that suggest that it is at least not obviously unfair. And I shall do so via a discussion of the history of Rawls's approach to the fairness condition over the course of the development of his theory.

4.1.5 Chapter 1 explained how, on the first form of Rawls's model, the two principles of justice appeared to be uniquely suited to satisfying the fairness condition. According to Rawls's theory at the time, this would be because they would be the principles that would be chosen in the original position. But I suggested that what made that choice fair was that they were the only principles that met the mutual advantage condition of Justice as Reciprocity.[1]

4.1.6 However, in Chapter 3, I maintained that the revised Justice as Fairness of the third model was actually unfit for the purpose of selecting principles that would meet the fairness condition. This was because the device of choice from behind a veil of ignorance made it possible for the parties to choose principles that didn't offer an improvement over the relevant situation of equal liberty. If a conception of justice doesn't meet the mutual advantage condition it can't meet the fairness condition, as it doesn't offer cooperating members any return for their cooperation it can't offer them a return that is fair.

4.1.7 This leaves open the question of what the fairness condition in the third model of the theory might involve. If only the two principles of justice could meet the mutual advantage condition (as was the case in Rawls's first model) then they would be the only conception of justice that could possibly meet the fairness condition. However, on the assumptions of the third model, all the conceptions of justice (bar general egoism) met the mutual advantage condition. But it does not follow from the stipulation that a conception of justice must meet the mutual advantage condition in order to count as fair, that all conceptions of justice that meet that condition are fair.

4.1.8 Although it is beyond the remit of this thesis to resolve the question of what the

[1] §§ 2.2.1 – 2.2.8

requirements of the fairness condition are, I am able to counter two arguments that might be thought to tell against the utilitarian conceptions of justice in favour of Rawls's two principles of justice. The first is that the two principles of justice are fairer because they are in some way neutral between conceptions of the good. The second is that the two principles of justice are fairer because they reject 'the justice of inequalities based on morally arbitrary advantages'.[1]

4.1.9 Chapter 2 argued that Rawls's first model of his theory was neutral between conceptions of the good (§§2.6.13 – 2.6.14). This was because the two principles of justice appeared to be the only conception of justice that could meet the mutual advantage condition of Justice as Reciprocity, and meeting the mutual advantage condition is a necessary condition for any conception of justice to be suited to the conception of Justice as Reciprocity. But the shift to the baseline of general egoism in the third model, allowing a variety of conceptions of justice to meet the mutual advantage condition, opens up the possibility of choosing between them on other grounds than that they meet the mutual advantage condition.

4.1.10 Rawls put forward some other considerations designed to argue that the two principles of justice were peculiarly suited to the conception of Justice as Reciprocity, revolving around the issue of stability; if those arguments were successful then there would be a case for claiming that selection of the two principles of justice was in some way neutral, rather than taking sides, between conceptions of the good. But in Chapter 3 I argued that the claim that the principles of justice were inherently more stable was a contingent matter (§§3.6.11 – 3.6.17). People might be more sympathetic to the worst off group, but they might be more sympathetic to helping intermediate groups or making the planet a better place for future generations. I also noted that the principle of utility could take care of any issues regarding stability internally (§3.6.18). The truth is that it is impossible to be neutral between conceptions of the good. All conceptions of the good will require sacrifices for others. The choice between the utilitarian conceptions of justice and the two principles of justice boils down to which 'others' those sacrifices should be made

[1] Barry 1989 p. 239

for. And that is a question about which is the preferable theory of benevolence. Strict neutrality is impossible.

4.1.11 The second claim is that the two principles of justice reject the injustice of inequalities based on morally arbitrary advantages. But, as Allan Gibbard has pointed out, they do not entirely reject the inequalities based on morally arbitrary advantages; the difference principle is prepared to bestow advantages on those who have natural talents but need incentives, for the sake of raising the level of the worst off group in society.[1] That constitutes inequalities based on morally arbitrary advantages. What *can* be said of the two principles of justice is that they don't set out to deliberately reward factors that are morally arbitrary from a moral point of view, as meritocratic theories of justice might do. But that could also be said of the utilitarian conceptions of justice.

4.1.12 In conclusion of this section, it appears that the demands of Justice as Reciprocity might turn out to be identical to the demands of Justice as Benevolence. All the conceptions of justice meet the mutual advantage condition. It is possible that for a lucky few, the demands of a particular conception of justice will exactly coincide with their conception of the good; in which case no constraint of behaviour will be required of those few.[2] Most, however, will have to constrain the pursuit of their conception of the good. The question of what the fairness condition might turn out to require has been left open; it may turn out to be unresolvable. There seems to be no reason within Justice as Reciprocity, on the showing so far, for a classical act-utilitarian to not believe that their doctrine is fully in conformity with Justice as Reciprocity.

2 From Justice as Reciprocity to Reciprocal Classical Utilitarianism

4.2.1 But the route from Justice as Reciprocity to classical act utilitarianism surely rests on one or more mistaken assumptions. Firstly, it is questionable whether the classical principle of utility will inevitably improve everyone's lot in comparison to a state of

[1] Gibbard p. 274

general egoism. As already noted, Nozick pointed out that the worst possible future state might be worse than 'the most pessimistically described state of nature'. It would be easy to devise science fiction scenarios where a minority are forced to endure prospects of not-worth-living lives for the sake of the good of advantages to the rest of the population. Secondly, it is debatable whether a state of general egoism is the relevant state of equal liberty by which to gauge mutual advantage. Mightn't it be possible for all to fare better than in a state of general egoism, but still have lives which weren't worth living? Does a life not worth living count as an 'advantage' to the person who lives it, even if they do enjoy better prospects than in the state of nature?

4.2.2 Fortunately, I don't need to provide definitive answers to these questions in order to show that people have the right to act in accordance with, and prescribe for others in society, the classical principle of utility as the predominant principle of distribution in society. The device I shall use to demonstrate that people have this right is to imagine the deliberations that Rufus T Firefly might go through if he were an 'ideal legislator' charged with the task of imposing institutions on Freedonia. But rather than being the utilitarian ideal legislator invoked by Rawls,[1] who was conceived of as solely concerned with arranging institutions in such a way as to maximize utility, he is to be conceived of as the Ideal Legislator (Reciprocity). As such, he must ensure that the institutions he imposes are suitable for the conception of Justice as Reciprocity, and incorporate whatever constraints on the pursuit of maximizing utility that conception of justice may require.

4.2.3 We can suppose, as I did in Chapter 2 (§2.6.14), that Firefly has utilitarian sympathies and knows that a substantial section of the population shares them. But he also knows that many of them have other conceptions of the good: some are prioritarians, some have strong religious convictions, and some are self-interested hedonists who may nevertheless have a sense of justice that would motivate them to follow the rules of society's governing institutions.

4.2.4 So Firefly's instinct would be to decree the classical principle of utility as Freedonia's official conception of the good. As asserted above, neutrality between

[1] Rawls 1999 p. 24

conceptions of the good is impossible (§5.1.10). But the fact that neutrality is impossible doesn't mean that all conceptions of the good are equally compatible with Justice as Reciprocity. The principle of classical utility is more liable to place people in a situation that is worse off absolutely than other conceptions of justice, such as the two principles of justice. Isn't that liability enough to disqualify it as suitable to Justice as Reciprocity?

4.2.5 My answer that it is not so disqualified, I shall argue, depends on Hume's doctrine of the circumstances of justice, which Rawls endorses.[1] Hume argues that in sufficiently adverse conditions 'the strict laws of justice are suspended…and give place to the stronger motives of necessity and self-preservation.'[2] He outlines several scenarios, including a shipwreck, in which he judges this suspension would be justified. Hume's interpretation of the doctrine of the circumstances of justice is, in part, a hedonistic one. He writes, '[t]he use and tendency of that virtue [i.e. the virtue of justice] is to procure happiness and security, by preserving order in society: but where the society is ready to perish from extreme necessity, no greater evil can be dreaded from violence and injustice; and every man may now provide for himself by all the means, which prudence can dictate, or humanity permit.'[3] For my argument that follows, I shall disregard Hume's reference to security, and assume that the circumstances under which there is no call for the exercise of the virtue of justice includes those circumstances where the exercise of that virtue would fail to procure sufficient happiness.

4.2.6 Now, the scenarios that Hume outlines are ones where all the people involved are in the same desperate boat. But the logic of his position, it seems to me, implies that the virtue of justice might be suspended for individuals or a section of society, if cooperation held out an insufficient chance of procuring happiness for them. In which case, if the classical principle of utility were to place some members of society in sufficiently adverse conditions, then they would no longer be in the circumstances of justice and, according to Hume's doctrine, the 'stronger motives of necessity and self-preservation' could take over from 'the strict laws of justice'.

---

[1] In *Theory* (*T of J Rev* 1999), Rawls remarks that 'Hume's account of them [i.e. the circumstances of justice] is especially perspicuous and the preceding summary adds nothing essential to his much fuller discussion.

[2] Hume 1902 p. 186

[3] Hume 1902 p. 186

4.2.7 The logic of Rawls's theory should commit him to the same position, for he writes that he 'shall, of course, assume that the persons in the original position know that these circumstances of justice obtain'.[1] If that is the case, then they should know that the people they represent are not in such dire straits that the laws of justice do not apply to them.

4.2.8 Rawls would have no doubt taken issue with my hedonistic reading of the circumstances of justice, but that turns out not to matter to the case that Rufus T Firefly has the right to impose the classical principle of utility on Freedonia.

4.2.9 We can imagine that Firefly does accept a hedonistic doctrine of the circumstances of justice. Furthermore, he accepts Sidgwick's view about rational self-interest.[2] So he believes that it would be self-interestedly rational for an individual to maximize their expectation of happiness over the whole course of their life. On Firefly's interpretation of the circumstances of justice, then, any Freedonian's obligation to act in accordance with the classical principle of utility would be suspended just when their overall life-expectancy of happiness became less than zero.

4.2.10 The circumstances of justice, thus interpreted, define a threshold which arguably must be met in order for people to have an obligation to comply with the classical principle of utility. The doctrine of the circumstances might be considered to define a basepoint, in addition to a situation of general egoism, which has to be exceeded in order for the mutual advantage condition of Justice as Reciprocity to be met. In which case, classical act-utilitarianism would not be compatible with Justice as Reciprocity. For classical act-utilitarianism requires everyone, no matter what position they might be in, to choose amongst those actions that would maximize utility.[3] There is no room in the theory to allow a suspension of the laws of justice.

4.2.11 But reciprocal classical utilitarianism can find room for such a suspension; it can simply relieve people from the obligation to act so as to maximize utility if doing so would mean that their own life expectation of happiness became less than zero. Suppose then, that Firefly instituted serfdom in Freedonia and it turned out that the position of serf was so

[1] Rawls *T of J Rev* 1999 p. 111
[2] Sidgwick pp. 89-95 and 119 - 122
[3] See Peter Singer's statement §1.0.4

miserable that if the serfs willingly complied with the duties of their station they would be very unlikely to have a life worth living. In their case the rules of justice would be suspended and they would have the right to rebel against Rufus or to attempt to flee to neighbouring Sylvania.

4.2.12 The right described so far is a liberty-right. Because the serfs would have this right, reciprocal classical utilitarian would uphold the core tenet of deontological liberalism, which grants certain people the liberty-right to not maximize utility.

4.2.13 Next, the question arises as to whether reciprocal classical utilitarianism shouldn't also accord people claim-rights. The agricultural workers worked up until the date when Firefly began to contemplate the imposition of serfdom; why aren't they owed the prospect of a worthwhile life, which would require others to respect at least their right to have that, and place them under the relevant constraints?

4.2.14 The answer to that question is that as soon as people lose the prospect of a life worth living they cease to count as persons who have a duty to cooperate, and hence as those to whom a duty of reciprocity is owed. The description of the requirements of Justice as Reciprocity in Chapter 1 confined the obligations to reciprocate to a closed community of benefactors and beneficiaries.[1] However, that does not mean that all the fruits of cooperation have to be confined to the closed community of benefactors and beneficiaries. If it is a community of utilitarians then they have the right to insist on rules that may

[1] In 'Constructing Justice' Gibbard appears to defend the right of a 'master' group to enslave a weaker group, asking 'What if a group can be enslaved without excuses, and enslaved profitably? The group is excluded from the terms of voluntary cooperation not because it has nothing to offer, but because it can be kept under control. Sufficient contribution can be exacted from members of the group without calling on their motives of fair reciprocity' (p. 272) He then goes on to speculate that '[i]f an extant fair scheme of social cooperation includes everyone, then everyone is owed fair reciprocity'. This is given as an interpretation of Rawls's underlying theory of Justice as Reciprocity and raises a couple of interesting questions about the overall viability of Justice as Reciprocity. If no extant schemes of social cooperation exist (as Barry asserts in *Theories of Justice*, also ascribing that view to Rawls (Barry 1989 p. 202)) then why do we not retain the right to be egoistic? In which case, why doesn't anyone have the right to try to enslave anyone else if they have the power to do so? Justice as Reciprocity threatens to collapse into Justice as General Egoism. I believe an answer may lie in the idea of people who have benefited enough from partially just societies (such as the one I live in) having obligations to make substantial sacrifices, but have not been able to develop it yet. In defence of the right to impose the principle of classical utility on the world, however, it can be observed that if we do have the right to do whatever we like in pursuit of our conception of the good then we certainly have the right to be a utilitarian.

distribute goods to third parties.

4.2.15 A second question arises as to whether society would have the right to impose a principle on society that requires all to maximize utility once they have reached the threshold required by reciprocal classical utilitarianism. Why shouldn't people receive benefits in proportion to contribution?

4.2.16 The trouble with the suggestion that contribution should be in proportion to benefit is that it appears to be impossible to gauge what any one individual's 'contribution' or 'benefit' would be when the relevant situation of equal liberty is a state of general egoism. A. J. Simmons has criticized the notion of using a state of general egoism as a yardstick to gauge benefit, asserting

> [i]t is far too easy to simply gesture at the horrors of a Hobbesian "war of all against all," concluding that of course all citizens benefit on balance from cooperative political schemes, standing as we do far above the level of well-being we would "enjoy" during a solitary, nasty, short life. The relevant baseline of comparison must include the effects of efforts at self-provision (or small group provision) of goods like security, efforts that would undoubtedly occur…in any realistic non-political situation.'[1]

But he offers no real justification for rejecting a Hobbesian war of all against all apart from the fact that it clashes with our intuitions about freedom and benefit. Simmons seems to be suggesting that some citizens might not be considered as benefiting from society because they could do better if they were in a small group outside society. But as Gibbard pointed out, 'a coalition that withdraws from society renounces any claim to justice from those who remain. Why think they could take along their per capita share of nonhuman productive assets? Why think they could avoid slavery at the hands of everyone else?'[2] Against a Hobbesian baseline the talented who are not allowed to emigrate, but enjoy a reasonable level of freedom within the society in which they are confined, do indeed count as benefiting from the restraint of others.

4.2.17 So far, I have suggested that the doctrine of the circumstances of justice may support a threshold which has to be met before the obligation to reciprocate kicks in, and

[1] Simmons 2001 p. 38
[2] Gibbard 1991 p. 272

given a hedonistic interpretation of what that threshold is. But that threshold and the hedonistic interpretation of it are both controversial claims. Many people alive today, I would speculate, live in societies where they have insufficient prospects of a happy life to qualify as having a duty of reciprocity on the account given so far but willingly comply with the rules of the societies they live in and believe they have an obligation to do so. Are they wrong?

4.2.18 They may be. But there are other thresholds that might be taken into account as alternatives. One is the state of general egoism which has already received plenty of attention.

4.2.19 The final threshold to be considered here is the definition of 'advantage' in the mutual advantage condition. I shall again consider this question within a hedonistic framework. If cooperation with the classical principle of utility affords a person insufficient chance of a happy life, does the society in question count as being to her advantage? This may turn out to be essentially the same as the threshold defined by the doctrine of the circumstances of justice.

4.2.20 There is, however, no need to provide definite answers to these questions surrounding the determination of the threshold relevant to reciprocal classical utilitarianism in order to sustain the claim that society has the right to embrace the classical principle of utility as the predominant principle of distribution for society. This is because whatever the answer turns out to be it will only provide those who fall beneath it a *liberty-right* to not maximize utility rather than a claim-right that places constraints on others' behaviour to respect that right. The point that no definite answers are needed to these questions can be clarified by considering the thought process that Firefly might go through before deciding whether to institute serfdom. He believes that the threshold is determined by a hedonistic interpretation of the circumstances of justice and also hopes, and expects, that serfdom will not be so bad as to place people below that threshold. But he can console himself with the thought that even if he turns out to be wrong on both counts he will still have the right to tie them to the land by force if necessary, and guard the borders to Sylvania to prevent people from leaving.

3 The three tenets of deontological liberalism

4.3.1 The main planks of the argument that reciprocal classical utilitarianism is only reconcilable with the core tenet of deontological liberalism have already been laid in section 2, with its claim that reciprocal classical utilitarianism would only require acknowledgement of a liberty-right. In this section I attempt to reinforce, and explain, the argument that Justice as Reciprocity does not underwrite the second tenet of deontological liberalism by first delving once more into the history of the development of Rawls's theory, and pointing to an analogy with Hobbes's theory in *Leviathan*. I then expand on my claim above that it is impossible for principles of justice to be neutral between conceptions of the good.[1]

4.3.2 In *Theory*, Rawls expressed his view that one of the key advantages of the contract view is its ability to underwrite our convictions of common sense regarding the nature of justice. He wrote

Passage 4a *T of J Rev*

> [1] It has seemed to many philosophers, and it appears to be supported by the convictions of common sense, that we distinguish as a matter of principle between the claims of liberty and right on the one hand and the desirability of increasing aggregate social welfare on the other; and that we give a certain priority, if not absolute weight, to the former. [2] Each member of society is thought to have an inviolability founded on justice or, as some say, on natural right, which even the welfare of every one else cannot override. [3] Justice denies that the loss of freedom for some is made right by a greater good shared by others. [4] The reasoning which balances the gains and losses of different persons as if they were one person is excluded. [5] Therefore in a just society the basic liberties are taken for granted and the rights secured by justice are not subject to political bargaining or to the calculus of social interests. [6] Justice as fairness attempts to account for these common sense convictions concerning the priority of justice by showing that they are the consequence of principles which would be chosen in the original position.[2] [my numbering of the sentences]

4.3.3 The 'claims of liberty and right' that Rawls maintains are 'supported by the convictions of common sense', given the context of *Theory*, in which Rawls maintains

[1] See §4.1.10
[2] Rawls *T of J Rev* pp. 24 - 25

that justice as fairness would result in the two principles of justice, should be interpreted as claim-rights; the basic liberties that are protected by the first principle of justice in *Theory* are certainly claim-rights, not liberty-rights. They should also, I suggest, be read as applying to real people in the real world; the common sense intuition that Rawls aims to support is that actual existing members of societies have rights, whether those societies respect them or not. If they are interpreted as claim-rights (as they should be), then Rawls's ambition in this passage should be read as hoping that the contractualist theory of justice as fairness will be able to account for the first two tenets of deontological liberalism.

4.3.4 Rawls goes on to argue that 'while the contract doctrine accepts our convictions about the priority of justice as on the whole sound, utilitarianism seeks to account for them as a socially useful illusion.'[1] This point can be exemplified by Mill's example of the 'qualified medical practitioner' in Chapter 1.[2] Mill believed very strongly that the kind of 'basic liberties' that would ordinarily protect a medical practitioner from being forced to do work she didn't want to do should be respected by society, perhaps almost to the extent of being regarded as inviolable. But ultimately he regarded the 'strength of these persuasions' as a 'socially useful illusion', as his being prepared to allow the kidnap of a medical practitioner in extreme circumstances attests.

4.3.5 It is true that utilitarianism historically has generally sought to account for the common sense precepts of justice corresponding to the first two tenets of deontological liberalism as socially useful illusions, and it is also true that classical act-utilitarianism, as reiterated throughout this thesis, because of its commitment to unlimited aggregation, can't account for them as sound.

4.3.6 But Rawls couldn't account for the first two tenets of deontological liberalism, I have argued in this thesis, because he never managed to solve the problem that it would be rational for anyone to prefer principles committed to unlimited aggregation in preference to a Hobbesian state of nature.

---

[1] Rawls *T of J Rev* pp. 24 - 25
[2] §1.5.5 and Passage 1e p. 39

4.3.7 My diagnosis of the fundamental problem can be supported by again looking into the history of the development of Rawls's theory. Rawls first put claims similar to those of **Passage 4a** in his essay 'Distributive Justice', when he was committed to the second form of the model. In 'Distributive Justice' he first put forward a claim equivalent to *Theory*'s claim that utilitarianism can only account for the common sense precepts of justice as a 'socially useful illusion', writing

Passage 4b *DJ* 1967

> From the standpoint of utility, the strictness of common-sense notions of justice has a certain usefulness, but as a philosophical doctrine it is irrational.
>
> If then, we believe that as a matter of principle each member of society has an inviolability founded on justice which even the welfare of everyone else cannot override, and that a loss of freedom for some is not made right by a greater sum of satisfactions enjoyed by many, we shall have to look for another account of justice…The aim of the contract doctrine is precisely to account for the strictness of justice by supposing that its principles arise from an agreement among free and independent persons in an original position of equality.[1]

4.3.8 I do not think any close comparison of these two passages is needed to justify reading **Passage 4b** as the historical antecedent of **Passage 4a**. In both these passages Rawls expresses his hope of accounting for the two tenets of deontological liberalism with Justice as Fairness.

4.3.9 However, the key ingredient which Rawls wanted to use to account for the two tenets of deontological liberalism in 'Distributive Justice' was the claim, closely scrutinized in Chapter 1, that rational individuals in the original position would not 'agree to a violation of their liberty for the sake of a greater balance of satisfactions enjoyed by others' but would 'insist upon institutions which returned compensating advantages for any sacrifices required.'[2] I argued in Chapter 1 that rational individuals would not reject a principle, such as the classical principle of utility, on the grounds that

[1] Rawls 1967 DJ p. 131
[2] Rawls 1967 DJ p. 132

it was committed to unlimited aggregation.[1] It would instead be rational for them to accept a principle committed to unlimited aggregation since it could still be *expected* to compensate all, including those who would end up worst off under it, for their 'sacrifice' of natural liberty, when that natural liberty was the Hobbesian liberty to pursue one's conception of the good by any means necessary. Rawls's hope of underwriting the first two tenets of deontological liberalism through justice as fairness, it seems to me, rested entirely on the notion that it would only be rational for the parties in the original position to choose non-aggregating principles over natural liberty.

4.3.10 My suspicion is that it is impossible to motivate claim-rights, and thereby underwrite the second tenet of deontological liberalism, through Justice as Reciprocity, without taking the relevant situation of equal liberty by which mutual advantage is to be gauged as one in which people already have claim-rights. Rawls's first model of his theory, as shown in Chapter 2, upheld the second tenet of deontological liberalism but took the relevant situation of equal liberty to be the equal liberty guaranteed by the first principle of justice. The model in between the first and second models, model 1.5, as we shall see below, also supposed that the relevant situation of equal liberty was a situation where people had claim-rights, though this time in the form of 'constitutional liberties' that the parties in the original position brought with them to the negotiating table. My reading of the second model, and the argument of 'Distributive Justice', is that Rawls hadn't fully managed to abandon the assumptions of the previous models and embrace the idea that rights would have to be justified as *compensation for a loss of natural liberty*, when that liberty is conceived of as the liberty of general egoism.[2]

---

[1] §§ 1.10.4 – 1.10.9

[2] My reading can be supported in two ways. First, the argument of Chapter 1 that examined the messy substitution in *Theory* of the claim that people wouldn't choose a principle that permitted sacrifices of the worse off for the better off, for the second model's premise that people wouldn't sacrifice their liberty for the sake of greater satisfactions to others but would insist on principles that gave them compensating advantages. Secondly, Rawls's case against the principle of utility in the original edition of Theory where he argues '[y]et should a person gamble with his liberties and substantive interests hoping that the application of the principle of utility might secure him a greater well-being, he may have difficulty abiding by his undertaking. He is bound to remind himself that he had the two principles of justice as an alternative.' (Rawls *T of J orig* 1971 p. 176). The 'people' referred to are the parties in the original position who are representative of real people occupying positions in society. But neither the parties in the original position, nor the real people in society should be conceived as already in possession of liberties. Whether they have liberties or not is for

4.3.11 But Justice as Reciprocity is able to underwrite the core tenet of deontological liberalism, as shown above (§§4.2.5-4.2.20), by insisting on a threshold below which people should not be allowed to fall. Those who do fall below it do not have an obligation to cooperate with the prevailing rules of society. The liberty right that those who fall beneath the relative threshold would retain is enough to account for the idea that 'each member of society [has] an inviolability founded on justice…which even the welfare of every one else cannot override.'[1] Reciprocal classical utilitarianism could accommodate this precept. It would say that a demand of classical act-utilitarianism for someone below this threshold to act so as to maximize utility would be overriding their liberty right to not maximize utility for the welfare of everyone else.

4.3.12 Here an analogy can be made with Hobbes's theory in *Leviathan* that helps illuminate the distinction between classical act-utilitarianism and reciprocal classical utilitarianism by first, supporting the idea that Justice as Reciprocity should acknowl9ge a threshold below which no-one should fall and secondly, supporting the idea that this threshold does not impact on society's right to embrace the classical principle of utility.

4.3.13 Hobbes famously argued that it was rational for all to surrender their natural liberty and agree to be unconditionally ruled by a Leviathan in exchange for everyone else doing the same. Nevertheless, he maintained that should the Leviathan attempt to kill any individual, that individual had the liberty-right to resist the Leviathan.[2] Similarly, Justice as Reciprocity holds that it would be rational for anyone to agree to a principle, such as the classical principle of utility, which is committed to unlimited aggregation as the preferable alternative to the relevant situation of equal liberty: a state of general egoism. Nevertheless, if the principle committed to unlimited aggregation threatened someone with an insufficient expectation of happiness, they would retain the right to not cooperate with the principle.

---

the parties in the original position to decide. This suggests to me that Rawls hadn't quite 'let go' of the assumption of his previous models. Rawls removed those two sentences for his second edition.

[1] In the words of **Passage 4a**.

[2] Hobbes 1996 pp. 145 -155

4.3.14 But Hobbes did not maintain that the right that some might have to resist the Leviathan led to a corresponding duty of the Leviathan to restrain his or her behaviour out of respect for that right. On the contrary, the Leviathan was perfectly free to attempt to kill whomsoever he or she pleased.

4.3.15 Similarly, a society which predominantly embraced the classical principle of utility would be free to apply that principle ruthlessly wherever it felt appropriate, even against those individuals who might be exempted from an obligation to comply with the principle themselves.

4.3.16 Chapter 2 maintained that the first model of Rawls's theory was better able to sustain the third tenet of deontological liberalism than subsequent models, though the Rawls of the first model was less obviously concerned with neutrality between conceptions of the good.[1] But he expressed a commitment to the third tenet of deontological liberalism clearly in *Justice as Fairness: A Restatement.*

Passage 4c: ***(J as F:AR 2001)***

> One role of political philosophy is to help us reach agreement on a political conception of justice, but it cannot show, clearly enough to gain general and free political agreement, that any single reasonable comprehensive doctrine, with its conception of the good, is superior. It does not follow (and justice as fairness as a political conception of justice does not say, and must not say) that there is no true comprehensive doctrine, or no best conception of the good. It only says that we cannot expect to reach a workable political agreement as to what it is. Since reasonable pluralism is viewed as a permanent condition of a democratic culture, we look for a conception of political justice that takes that plurality as given.[2]

4.3.17 There is an implicit claim underlying this passage that Rawls's conception of justice – the two principles of justice – is neutral. I have argued that its appearance of neutrality is deceptive; accepting the difference principle favours those that would prefer to prioritize the worst off rather than promoting the greatest

[1] §§2.6.12 – 2.6.16

[2] Rawls 2001 p. 84

happiness of the greatest number. The first principle of justice has more claim to neutrality, but even that can be, and has been, accused of promoting a liberal conception of the good in preference to other conceptions. The reason that the first model was better able to sustain the third tenet of deontological liberalism was that there appeared to be no room for considerations of what might be the best conception of benevolence to operate; meeting the demands of the mutual advantage condition mandated the second principle of justice. The argument of this chapter, and the two preceding it, has, I hope, shown that a variety of conceptions of justice would meet the mutual advantage condition, including all the utilitarian conceptions of justice. Where, in the first model, the worry was that any aggregation would violate the mutual advantage condition, the argument of this chapter should have shown that Justice as Reciprocity need not be concerned with that at all. The unlimited aggregation that the classical principle of utility embraces is compatible with Justice as Reciprocity as, if combined with the threshold required by reciprocal classical utilitarianism, the worst unlimited aggregation can do to people is grant some of them the liberty-right to not maximize utility. It can't violate anyone's rights.

4.3.18 The way is open, then, for society to choose which principles of justice to embrace on the grounds of which best fulfils the requirements of benevolence. Interestingly, Rawls accepts in **Passage 4c** that there may be 'a true comprehensive doctrine' and 'best conception of the good'. If utilitarians can win the case for the classical principle of utility being the best conception of the good, then they are entitled to try to promote it without fear that they will be acting unjustly according to the conception of Justice as Reciprocity.

4.3.19 Of course, this does not mean that utilitarians must reject liberalism, or the 'reasonable pluralism' that Rawls was so concerned to accommodate, or that ISUS[1] should try to emulate the example of ISIS[2] and carve out a consequentialist caliphate across the Western world. If Mill was right, then utilitarianism may

[1] The International Society of Utilitarian Studies
[2] The Islamic State of Iraq and Syria

provide a solid foundation for the liberal values that Rawls was so concerned to uphold. And if my arguments are correct, the purported foundations for them in Justice as Reciprocity are very shaky indeed.

4 The separateness of persons

4.4.1. Possibly Rawls's most influential criticism of utilitarianism in *A Theory of Justice* is his charge that it ignores 'the separateness of persons'.[1] This criticism of utilitarianism is accepted by some who have little regard for other aspects of Rawls's theory, such as his contractualism and his two principles of justice. And amongst those who are at least somewhat sympathetic toward Rawls's philosophy, there is no clear consensus on how the separateness of persons' objection to utilitarianism should be interpreted. Some commentators, following Derek Parfit, have thought that we should prioritize benefits to the worst off in society in order to meet the objection.[2] Others have claimed that the separateness of persons constitutes an objection to the priority view.[3] Many commentators have associated the separateness of persons with egalitarianism, though it has been argued that it counts against 'luck-egalitarianism'.[4]

4.4.2 The broadest area of consensus appears to be that taking the separateness of persons seriously will result in a repudiation of utilitarian aggregation. And some who hold this position also argue that taking the separateness of persons seriously motivates claim-rights.[5] But it has been suggested that this would not necessarily be so if utilitarianism was derived through contractualism.[6]

---

[1] William Shaw remarks that '[i]n the past twenty-five years, many philosophers have been persuaded by John Rawls that the root problem is that utilitarianism ignores "the separateness of persons." So widespread is this contention that it has become a virtual mantra.'
[2] For example, Thomas Porter in 'In Defence of the Priority View' (2012) and Martin O'Neill in 'Priority, Preference and Value' (2012).
[3] Otsuka and Voorhoeve 2009, Otsuka 2012
[4] Voorhoeve and Fleurbaey 2012
[5] E.g. Vallentyne 2006 and Nozick 1974
[6] Mckerlie 1988 and Hirose 2013

4.4.3 The contractualist interpretation is, in my opinion, along the right lines. But I believe that the commentators who endorse the contractualist interpretation have missed something important by focussing on the objection as it was put in *Theory*. In my view, the separateness of persons' objection can only be properly understood in the context of the essay in which it was originally formulated, Rawls's 1963 essay 'Constitutional Liberty and the Concept of Justice.'

4.4.4 Here I put the case that the separateness of persons' objection is an objection that is, at its core, very similar to the 'exclusion of aggregation' considered in Chapter 1. In its original version it can, at least partly, be read as an objection that utilitarian aggregation would violate the mutual advantage condition of Justice as Reciprocity. I shall also argue that the objection that utilitarian aggregation would violate the mutual advantage condition is the most important part of the objection from the perspective of Justice as Reciprocity. Finally, I suggest that reciprocal classical utilitarianism would take the separateness of persons' seriously where classical act utilitarianism wouldn't.

4.4.5 In order to understand the separateness of persons' objection properly, we need to first understand the state of development Rawls's theory was in at the time he first put the objection. That state lay halfway between Wolff's 'first and second models', so I shall call it **model 1.5.** The important similarities to the first model were first; that both principles were still supposed to apply to 'liberty', understood as an assignment of rights to different positions in the practices of society, and secondly; that the second principle of justice conferred its 'advantageous inequalities' to the relevant people in the form of greater 'liberty', also understood as an assignment of such rights.[1]

4.4.6 But there were two highly significant modifications to Rawls's first model. One was that Rawls had redefined the first principle in terms of the protection of a specific, and limited, set of 'constitutional liberties'. This modification would be retained in *Theory* where the set of liberties would be slightly altered and renamed the 'basic liberties'. The

[1] This reading of model 1.5 is justified by the wording of Rawls's explanation of how the inequalities permitted by the second principle of justice should be understood in CL&CJ p. 75, which is virtually identical to the wording of Passage 2c p. 82

second was that Rawls explicitly equated the liberty of the original position with the liberty that would be secured by the first principle of justice. These two modifications are revealed by the following two passages

Passage 4d (*CL & CJ 1963)*

> [1][t]o satisfy the concept of justice, there must then exist in society a position of equal citizenship within which the liberty of the person is secured and which will express in institutions *the satisfaction of the first principle*. [2] Given this equal liberty, there will exist a position from which the application of the second principle of justice may be discussed and a sense of community maintained.[1] [my italics and numbering of the sentences]

Passage 4e (*CL & CJ 1963)*

> [1] It is now possible to comment on why it is characteristic of *the constitutional liberties* that they must be equal; why, in respect to these liberties as defined by the structure of the social system in which each begins, no one can be favoured. [2] Their role is to mark off and to define a part of the social structure distinct from that part which allows differences in rights and powers and a varied distribution of good things in accordance with the second principle. [3] Roughly, the distinction between these two parts of the social structure is that the first part – the constitutional liberties – expresses in institutions *the original position of equal liberty*; *it represents the position from which the application of the second principle may proceed among persons secure in their fundamental equality.* [4] The second part of the social structure contains those distinctions and of political, economic, and social forms necessary for efficient and mutually beneficial arrangement of joint activities; but such distinctions can be acknowledged only in matters of secular and personal interests or, roughly speaking, in matters of welfare.[2] [my italics and numbering of the sentences]

4.4.7 The reading of the two aforementioned modifications into the passages can be justified as follows. Sentence [1] of **Passage 4d (*CL & CJ 1963)*** describes the first principle of justice as securing a position of equal citizenship 'within which the liberty of the person is secured'. Sentence [2] of **Passage 4e (*CL & CJ 1963)*** clarifies that the

[1] Rawls 1963 p. 84
[2] Rawls 1963 p. 88

'liberty of the person' secured by the first principle is the person's possession of the constitutional liberties. Sentence [3] equates the position of equal liberty with the liberty of the original position.

4.4.8 Now it should be noted that this model really is quite different to both the first model of ***J as F (1&2***) and the third model of *Theory*. The differences to the first model have already been discussed. The two significant differences to the third model are first - that the people in the original position know their identities and how the various ways the second principle of justice might be applied would affect them; and second, that they arrive at the negotiating table already in possession of their rights or liberties. These rights should certainly be understood as meeting the requirements of the second tenet of deontological liberalism, which imposes on people a duty to restrain their behaviour in order to respect the rights of others: in other words, which accords people 'claim-rights'. Rawls's extended discussion of the question of slavery in ***CL&CJ*** makes it quite clear that he did not only believe that slavery would violate a person's right to be free, but that others had to respect that right.[1]

4.4.9 With the background in place I can put forward my argument that the separateness of persons' objection should be understood primarily as an objection to aggregation that would violate the mutual advantage condition. There are, I believe, three distinct separateness of persons' objections that need to be distinguished: the 'exclusion of aggregation', 'the unanimity objection' and 'the construction objection'.

4.4.10 The first two of these are contained in the following passage

Passage 4f (*CL & CJ* 1963)

> [A] The peculiar feature of the concept of justice is that it treats each person as an equal sovereign as it were and requires a *unanimous acknowledgement*

[1] See CL & CJ pp. 82-85. It should be recalled that the if people only had a right to 'freedom of the person' interpreted as a liberty-right, then no-one could be under the obligation to accept the bondage of slavery, but no-one would be under the obligation not to force slavery on people if they had the power to do so.

> from an *original position of equal liberty.* [B] In contrast to the conception of social utility, the concept of justice *excludes the possibility of arguing that the violation of the claims of some is justified (rendered just) by compensating advantages to others.* [C] If there is any disadvantage which cannot be acknowledged, an institution is unjust.[1] [my sentence lettering and italics]

4.4.11 As we saw above, in **Passage 4a *T of J Rev*[2],** Rawls wrote that '[3] Justice denies that the loss of freedom for some is made right by a greater good shared by others.'[3] If we bear in mind that what the parties in the original position have claim to in Model 1.5 *includes their constitutional liberties*, it becomes apparent that Sentence [B] is ***CL & CJ***'s equivalent to *Theory*'s Sentence [3]. Slightly more explanation is needed to show that it is closely related to ***J as F (2)***'s 'exclusion of aggregation.'

4.4.12 The exclusion of aggregation was expressed in **Passage 2g (*J as F 2*)**[4] Sentence [4]: 'The principle excludes, therefore, the justification of inequalities on the grounds that the disadvantages of those in one position are outweighed by the greater advantages of those in another position.'[5] At first sight this may appear to be a completely different objection. But it should be recalled that in Rawls's first model, 'disadvantages', 'advantages', and 'inequalities' were all functions of the assignment of rights - i.e.'liberties'- that different citizens had. The justification of any inequalities on the grounds that the disadvantages of those in one position were outweighed by the greater advantages of those in another position would, then, deprive the disadvantaged citizens of the maximum 'equal liberty' they were guaranteed by the two principles of justice. So in the first model, such justification could have been described in the terms used in Sentence [B] as 'justifying the violation of the claims of some (to the most extensive liberty compatible with a like liberty for all – measured in terms of economic advantage) by compensating advantages to others.'

4.4.13 The unanimity objection is expressed in both sentences [A] and [C]. The implication is that the kind of inequalities that utilitarian aggregation would be liable to

---

[1] Rawls 1963 p. 95
[2] p. 185
[3] § 4.3.2
[4] p. 88
[5] See §2.1.17.

lead to would not receive unanimous acknowledgement in the original position.

4.4.14 The construction objection is expressed in the paragraph following **Passage 4f (*CL & CJ* 1963).**

Passage 4g (*CL & CJ* 1963)

> [1] The concept of justice is distinct from that of social utility in that justice takes the plurality of persons as fundamental, whereas the notion of social utility does not. [2] The latter seeks to maximize one thing, it being indifferent in which way it is shared among persons except insofar as it affects this one thing itself. [3] *The conception of utility extends the principle of rational choice for one person to the case where there is a plurality of persons, for one person may properly count his advantages now as compensating for his own losses earlier or subsequently, but justice excludes the analogous reasoning between persons.* [4] The plurality of persons must construct among themselves the principles in accordance with which they are to decide between institutions, and the general characterization of the circumstances of this construction and the respective positions of persons within it are given by the analytic framework by which the two principles of justice were derived. [my sentence numbering and italics]

4.4.15 Sentence [3] is Rawls's first description of **the utilitarian conception of society**, described at length in Chapter 1 (§§1.8.1 – 1.8.4). The construction objection can be defined as the objection that extending the principle of rational choice for one person to the case of society is the wrong way to go about constructing principles of justice.

4.4.16 So there appear to be three distinct objections involved in 'the separateness of persons' objection to utilitarianism. I can put my case that the most important objection is the exclusion of aggregation by again using the example of Rufus T Firefly, conceived of as the Ideal Legislator (Reciprocity) contemplating whether to impose serfdom on the agricultural workers of Freedonia.

4.4.17 So far we have not considered how Firefly may have derived his utilitarian sympathies. But let us now suppose that he embraces the utilitarian conception of society, so he derives his utilitarianism by applying the principle of rational choice for one person to society. In making this decision he has treated separate lives as if they were separate

moments in the same person's life, which is enough to make his decision fall foul of the construction objection. But that still leaves the question as to exactly what is objectionable about making such decisions open. This point can be brought home by observing that Sentence [3] of **Passage 4g** (CL & CJ 1963) consists of a descriptive claim in its first two clauses ('The conception of utility extends the principle of rational choice for one person to the case where there is a plurality of persons, for one person may properly count his advantages now as compensating for his own losses earlier or subsequently') and a normative one in its third clause ('justice excludes the analogous reasoning between persons'). Firefly's decision fits the description of the first two clauses, and so is subject to the construction objection. But why is that unjust? And a second, related, question also naturally arises: what would 'taking the separateness of persons seriously' involve?

4.4.18 These questions are not clearly answered in *Theory*, as the varying interpretations of the separateness of persons' objection attests. By contrast, I would maintain that Rawls answered them quite clearly in 'Constitutional Liberty and the Concept of Justice'. Sentence [1] of **Passage 4g** stipulates that justice**,** in contrast to social utility, takes the plurality of persons as fundamental and Sentence [4] goes into more detail about how that should be done: the plurality of persons – i.e. the persons in the original position – must construct the principles themselves. **Passage 4f**[1] goes into still more detail in Sentences [A] and [C]: the principles the persons in the original position will construct will be those that can be unanimously acknowledged. The principle of utility, by implication, cannot be unanimously acknowledged. So the Rawls of 'Constitutional Liberty and the Concept of Justice' appears to have regarded the unanimity objection as the core objection. In answer to the question of why it is unjust to apply the principle of rational choice for one man to society, Rawls would have answered that doing so would be liable to lead to disadvantages which can't be acknowledged. And in answer to the question of what taking the separateness of persons seriously involves, he would have answered that it involves coming up with principles that could be unanimously acknowledged in the original position.

4.4.19 But, with arguments similar to the ones made in Chapters 2 and 3, I would

[1] pp. 193 - 194

maintain that the exclusion of aggregation is the core objection.[1] In Model 1.5, as in the first model, the two principles of justice are the only conception of justice that offer an improvement over the relevant situation of equal liberty by which to measure mutual advantage. In Model 1.5, the relevant situation of equal liberty by which to measure mutual advantage, as explained above (§§4.4.6 – 4.4.8), was a situation in which all are already assured of their constitutional liberties. A principle committed to unlimited aggregation, then, would be liable to violate people's liberties without giving them compensation for that violation. If Rufus T Firefly imposed serfdom on Freedonia, serfdom would presumably violate one or other of the agricultural workers' constitutional liberties, 'liberty of the person' and/or 'freedom of movement'.[2] And if Firefly did this for the sake of greater advantages to the non-agricultural workers, it could be maintained that this violation of the constitutional liberties was justified (made just) by compensating advantages to the non-agricultural workers. This would be unjust according to the exclusion of aggregation expressed in Sentence [B] of **Passage 4f[3].** In depriving them of this liberty it would also place the agricultural workers below the relevant situation of equal liberty and thereby violate the mutual advantage condition of Justice as Reciprocity.[4]

---

[1] §§2.2.8 – 2.2.10 and §§3.1.4 – 3.1.5

[2] Rawls 1963 p.74. These are among the constitutional liberties listed by Rawls, but he does not describe their content.

[3] pp. 193 – 194.

[4] It is useful, I think, to locate my interpretation of the 'core' separateness of persons' objection in relation to other interpretations. Larry Temkin (2015 pp. 98 – 99) interprets Rawls as taking the construction objection to be central. Iwao Hirose (Hirose 2013 p.193) reads the exclusion of aggregation to be the Rawls's main 'separateness of persons' objection. Interestingly, McKerlie (1988 p. 209) reads the exclusion of aggregation (which he refers to, following Parfit, as 'the objection to balancing') to be based, not on the separateness of persons, but on 'a rights view assigning certain basic liberties to everyone'. McKerlie's view is closest to my own by taking Rawls's insistence that just principles must take the separateness of persons' seriously as a demand for contractualism. He then makes the point that taking the separateness of persons' seriously would only motivate egalitarianism, if egalitarianism would be the result of contractualism. If contractualism led to utilitarianism then utilitarianism would take the separateness of persons' seriously, rather than egalitarianism (McKerlie 1988 pp. 214-215). However, McKerlie just focusses on the objection as it is presented in *Theory*. In *CL&CJ*, contractualism would decisively reject the principle of utility as those who would be 'losers' under the principle of utility would know their identities and have the opportunity to veto it. I have referred to the contractualist objection in *CL&CJ* as the 'unanimity objection' to emphasize the point that in *CL&CJ*, Rawls's contract required unanimity amongst people who knew their identities and how various principles

4.4.20 Now it should be easy to understand how the 'exclusion of aggregation' might be regarded as the core separateness of persons' objection, and the other two are of lesser importance. Consider first the construction objection. Applying the principle of rational choice for one man to the case of society is obviously not the right way to go about constructing principles suited for Justice as Reciprocity as it takes no account of the mutual advantage condition of Justice as Reciprocity. But suppose Firefly did not derive his utilitarianism by applying the principle of rational choice for one man to society, but had another foundation for his utilitarianism.[1] His utilitarianism would still lead him to impose serfdom on Freedonia and violate the mutual advantage condition. Consider next the unanimity objection. In model 1.5 the agricultural workers, who know how the principle of utility would affect them, would not acknowledge the disadvantages imposed on them by serfdom, so utilitarian aggregation would be subject to the unanimity objection. But the reason they would not acknowledge these disadvantages is because they disadvantage them with respect to the relevant situation of equal liberty, i.e. they violate the mutual advantage condition.

4.4.21 My interpretation of Rawls's core separateness of persons' objection to utilitarianism as presented in CL&CJ can be summarized, then, as an objection that utilitarianism would be liable to deprive some people of what they would have claim to in the relevant situation of equal liberty and to justify that deprivation in terms of compensating advantages to others. Importantly, for the forthcoming reconsideration of whether classical utilitarianism would be subject to the objection given the revised assumptions of *Theory*, it should be noted that this interpretation does not necessarily object to the violation of the claims of some so long as they receive compensation for that violation *themselves*. Classical act utilitarianism and

---

would be likely to affect them. The contractualism of *CL&CJ* can lay fair claim to take the separateness of persons more seriously than the contractualism of *Theory*. However, for the reasons laid out below (§4.4.20), I would maintain that the exclusion of aggregation still represented the core objection in *CL&CJ*.

[1] In *Reasons and Persons*, Derek Parfit suggested that a 'reductionist view of personal identity' provides a sounder foundation for utilitarianism (Parfit 1984 pp. 330 – 345)

reciprocal classical utilitarianism would both be subject to the objection on this interpretation, in model 1.5, as they would both be liable to justify the violation of the claim rights people have in the relevant situation of equal liberty (the original position) by compensating advantages to *others*.[1]

4.4.22 However, the revised assumptions of *Theory* reopen the question of whether utilitarianism would be subject to the objection in the third model. As explained in Chapter 1, the relevant situation of equal liberty in *Theory* by which mutual advantage is to be gauged is a state of general egoism, and all the conceptions of justice, including the utilitarian ones, were assumed to be advantageous to all in comparison with this benchmark.[2] In which case, if Rufus T Firefly imposed classical act-utilitarianism and serfdom on Freedonia, he could try to argue that he was not violating *Theory*'s exclusion of aggregation as expressed in **Passage 4a**[3] Sentence [3]. He was not making right the loss of freedom for the agricultural workers by a greater good shared by the non-agricultural workers. The agricultural workers receive compensation for their loss of natural liberty – understood as the Hobbesian liberty to do whatever they wanted in pursuit of the good – themselves. This compensation came in the form of a greater expectation of happiness than they would enjoy had they retained their natural liberty.

4.4.23 That line of argument is problematic, for the same reasons that the argument from Justice as Reciprocity to Classical Act-Utilitarianism discussed above was problematic.[4] It is questionable whether classical utilitarianism would improve everyone's position with respect to a state of general egoism or, even if it did, whether lives with sufficiently dismal expectations should count as advantageous to those who have such low life expectations. To illustrate this point with reference to the example in hand, if serfdom turned out to be sufficiently miserable, then any assertion by Firefly that their loss of natural liberty was made right by compensating advantages to the serfs – even when that liberty is understood as the liberty of

[1] As demonstrated with the example of Firefly and Freedonia above (§4.4.20)
[2] §§1.9.32 – 1.9.40
[3] p. 185
[4] §§ 4.1.1 – 4.1.12

general egoism for all – would be risible. Classical act utilitarianism would be subject to the separateness of persons' objection because it would deprive some people of their natural liberty without giving them adequate compensation in return. Because they received insufficient compensation for their loss of natural liberty *themselves*, they could be described as having the violation of their claim to what they would have in the relevant situation of equal liberty – the right to pursue their own conception of the good by any means necessary – justified by compensating advantages to *others*.

4.4.24 But reciprocal classical utilitarianism would not fall foul of my interpretation of the core separateness of persons' objection. Those who were offered sufficiently dismal life expectations by the classical principle of utility would retain their liberty-right to pursue their conception of the good by any means necessary.[1] So no one would be in the position of having the violation of their claim to what they would have in the relevant situation of equal liberty – the right to pursue their own conception of the good by any means necessary – justified by compensating advantages to others. Either they would receive compensation for their loss of natural liberty themselves, or they would retain their natural liberty. It is for this reason that I emphasized that my interpretation of the core separateness of persons' objection did not necessarily object to the violations of the claims of some so long as they received compensation for that violation themselves.

4.4.25 Reciprocal classical utilitarianism could also be described as taking the separateness of persons' seriously in a way that classical act-utilitarianism wouldn't. Sentence [2] of **Passage 4g (*CL & CJ* 1963)** gives a description of how 'the notion of social utility' might be regarded as not taking the separateness of persons' seriously: it 'seeks to maximize one thing, it being indifferent in which way it is shared among persons except insofar as it affects this one thing itself.' This theme was retained in *Theory* where Rawls wrote

> On this [the utilitarian] conception of society separate individuals are

[1] The question of what threshold is relevant for Justice as Reciprocity is discussed above §§ 4.2.1 – 4.2.20

> thought of as so many different lines along which rights and duties are to be assigned and scarce means of satisfaction allocated in accordance with rules so as to give the greatest fulfillment of wants. The nature of the decision made by the ideal legislator is not, therefore, materially different from that of an entrepreneur deciding how to maximize his profit by producing this or that commodity, or that of a consumer deciding how to maximize his satisfaction by the purchase of this or that collection of goods.[1]

4.4.26 The ideal legislator in this passage is the ideal legislator Rawls associated with the concept of Justice as Benevolence.[2] The Ideal Legislator (Benevolence) is to take the maximization of utility to be the 'one thing' that determines the distribution of rights and duties. But this is not true of the Ideal Legislator (Reciprocity). As argued above (§§4.2.2 – 4.2.20), he or she would have to take a threshold into account, and those citizens who would fall below the threshold if they acted in accordance with the classical principle of utility would retain the right to pursue their conception of the good by whatever means necessary. These rights would not be determined by the maximization of utility; it is quite conceivable that maximizing utility would assign those with very dismal life prospects a duty to act so as to maximize utility rather than a right to pursue their conception of the good. So, in contrast to classical act-utilitarianism, reciprocal classical utilitarianism would be concerned with the distribution of two things: utility and 'natural' rights. The rights can be appropriately referred to as natural, as they are the rights that people would have in the 'state of nature', understood as a state of general egoism. So reciprocal classical utilitarianism could be described as taking the separateness of persons seriously in the terms of **Passage 4g (*CL & CJ* 1963):** it would not 'be indifferent in which way..[one thing] is shared among persons except insofar as it affects this one thing itself.'

4.4.27 An important implication of my interpretation of the separateness of persons' objection to utilitarianism is that 'taking the separateness of persons seriously' would not, as some commentators have supposed, establish claim-rights

[1] Rawls 1971 p. 27. In fact, as seen in Chapter 2 of this thesis (§§2.4.12 – 2.4.19), the idea of the ideal legislator organising rights and duties with the sole aim of giving 'the greatest fulfilment of wants' appeared in ***J as F*** (1) so predated Rawls's utilitarian conception of society.
[2] See §§ 1.8.5 – 1.8.8 and §§2.4.12 – 2.4.19

but only liberty- rights.[1] This is because, as explained above (§4.4.24), the right that those who would not receive sufficient compensation for their loss of natural liberty through the application of the classical principle of utility would retain – the right to act egoistically – is a liberty-right, not a claim right. It does not place anyone under a duty to constrain their own behaviour out of respect for that liberty right.

4.4.28 In summary of this lengthy section, I have offered an interpretation of the separateness of persons' objection that upholds it as a genuine objection to classical act utilitarianism but not to reciprocal classical utilitarianism. It does not establish the injustice of embracing the classical principle of utility as the predominant distributive principle for society. And my interpretation has located the objection firmly within Rawls's conception of Justice as Reciprocity: its core is best understood as an objection to violations of the mutual advantage condition. So it is not, as some have supposed, an objection to utilitarianism that is independent of Rawls's conception of Justice as Reciprocity.

4.4.29 In conclusion of this final chapter and the thesis, I hope to have shown that Justice as Reciprocity is at least reconcilable with the classical principle of utility. This should be a welcome result for those who, as I described my own position in Chapter 1[2], were attracted to the classical principle of utility as a principle of benevolence but were concerned about its justice.

[1] As already remarked, (Footnote 4 above) Nozick (1974) and Vallentyne (2006) are two philosophers who maintain that taking the separateness of persons' objection to utilitarianism motivates claim-rights.
[2] §§1.0.1 – 1.0.5

Bibliography

Barry, B. (1989). *Theories of Justice.* Los Angeles: University of Californa Press.

Barry, B. (1995). *Justice as Impartiality.* Oxford: Clarendon Press.

Brink, D. (1993). 'The Separateness of Persons, Distributive Norms and Moral Theory'. In R. G. Frey, *Value, Welfare and Morality* (pp. 252 - 289). Cambridge: Cambridge University Press.

Broome, J. (1991). *Weighing Goods.* Oxford: Clarendon.

Cohen, J. (1989). 'Democratic Equality.' *Ethics 99(4)* 727 - 751

Curran, E. (2010). 'Blinded by the Light of Hohfeld: Hobbes's Notion of Liberty.' *Jurisprudence 1(1)* 85-104.

Driver, J. (2012). *Consequentialism.* Abingdon: Routledge.

Dworkin, R. (1978). 'Justice and Rights.' In R. Dworkin, *Taking Rights Seriously* (pp. 150 – 184). Cambridge: Harvard University Press.

Freeman, S. (2007). *Rawls.* Abingdon: Routledge.

Gauthier, D. (1985). *Morals by Agreement.* Oxford: Clarendon Press.

Gibbard, A. (1991). 'Constructing Justice.' *Philosophy & Public Affairs 20(3)* 264-277.

Gibbard, A. (2008). *Reconciling Our Aims.* Oxford: Oxford University Press.

Goldman, H. S. (1980). 'Rawls and Utilitarianism.' In G. B. Smith, *John Rawls' Theory of Social Justice* (pp. 346 -384). Ohio: Ohio University Press.

Hirose, I. (2013) 'Aggregation and the Separateness of Persons.' *Utilitas 25(2)* 182 - 195

Hobbes, T. (1996). *Leviathan.* Cambridge: Cambridge University Press.

Hume, D. (1902). *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*. Oxford: Clarendon.

Locke, J. (2003). *Two Treatises of Government and A Letter Concerning Toleration.* New Haven: Yale University Press.

Mckerlie, D. (1988). 'Egalitarianism and the Separateness of Persons.' *Canadian Journal of Philosophy 18(2)* 205 – 225.

Mill, J. S. (2003). *Utilitarianism and On Liberty.* Oxford: Blackwell.

Nagel, T. (1970). *The Possibility of Altruism.* Oxford: Clarendon.

Nagel, T. (1975). 'Rawls and Justice.' In N. Daniels, *Reading Rawls* (pp. 1 - 16). Oxford: Basil Blackwell.

Norcross, A. (2009). 'Two Dogmas of Deontology' *Social Philosophy & Policy 26(1)* 76 – 95.

Nozick, R. (1974). *Anarchy, State and Utopia.* Oxford: Blackwell.

O' Neill, M (2012). 'Priority, Preference and Value.' *Utilitas 24(3)* 332 – 348.

Otsuka, M (2012). 'Prioritarianism and the Separateness of Persons'. *Utilitas 24(3)* 365 – 380.

Otsuka, M and Voorhoeve, A (2009). 'Why it Matters that Some are Worse Off than Others: An Argument Against the Priority View'*, Philosophy & Public Affairs 37(2)* 169–97.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Clarendon.

Parfit, D. (2002) 'Equality and Priority' In M. Clayton and A. Williams, *The Ideal of Equality* (pp. 81 – 126) Basingstoke: MacMillan.

Porter, T. (2012). 'In Defence of the Priority View.' *Utilitas 24(3)* 349 – 364.

Rawls, J. (1955).' Two Concepts of Rules.' *The Philosophical Review 64(1)* 3-32.

Rawls, J. (1957). 'Justice as Fairness.' *The Journal of Philosophy 54(22)* 653-662.

Rawls, J. (1958). 'Justice as Fairness.' *The Philosophical Review 67(2)* 164-194.

Rawls, J. (1963). 'The Sense of Justice.' *The Philosophical Review 72(3)* 281-305.

Rawls, J. (1967). 'Distributive Justice'. In Freeman, S. (Ed.) (1999) *John Rawls: Collected Papers* (pp. 130 - 154). Cambridge: Harvard University Press.

Rawls, J. (1968). 'Distributive Justice: Some Addenda.' *Natural Law Forum*, 51-71.

Rawls, J. (1971). *A Theory of Justice (Original Edition).* Cambridge: Belknap.

Rawls, J. (1985). 'Justice as Fairness: Political not Metaphysical.' *Philosophy & Public Affairs 14(3)* 223 - 251.

Rawls, J. (1996). *Political Liberalism.* New York: Columbia University Press.

Rawls, J. (1999). *A Theory of Justice (Revised Edition).* Cambridge: Belknap.

Rawls, J. (2001). *Justice as Fairness: A Restatement.* Cambridge: Belknap.

Scanlon, T. (1982). 'Contractualism and Utilitarianism.' In A. Sen and B. Williams, *Utilitarianism and Beyond* (pp. 103 - 129). Cambridge: Cambridge University Press.

Scheffler, S. (1994). *The Rejection of Consequentialism.* Oxford: Clarendon Press.

Scheffler, S. (2003). 'Rawls and Utilitarianism.' In S. Freeman, *The Cambridge Companion to Rawls* (pp. 426 - 459). Cambridge: Cambridge University Press.

Shaw, W. (1999) *Contemporary Ethics: Taking Account of Utilitarianism*. Oxford: Blackwell

Sidgwick, H. (1907). *The Methods of Ethics*. London: MacMillan

Simmons, A. J. (2001) 'Fair Play and Political Obligation: Twenty Years Later'. In A. J. Simmons, *Justification and Legitimacy: Essays on Rights and Obligations.* Cambridge: Cambridge University Press.

Singer, P. (1981). *The Ever Expanding Circle.* Princeton: Princeton University Press.

Smith, A. (2007). *The Wealth of Nations.* Hampshire: Harriman House.

Temkin, L. S. (2012) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning.*

Vallentyne, P. (2006). 'Against Maximizing Act-Consequentialism.' In J Dreier, *Contemporary Debates in Moral Theory* Oxford: Blackwell.

Voorhoeve, A. and Fleurbaey, M. (2012). 'Egalitarianism and the Separateness of Persons', *Utilitas 24(3)* 381 - 398.

Williams, B. (1973). 'A critique of utilitarianism.' In B. J. Williams, *Utilitarianism: For and Against* (pp. 77 - 151). Cambridge: Cambridge University Press.

West, H. R (2004). *An Introduction to Mill's Utilitarian Ethics.* Cambridge: Cambridge University Press

Wolff, R. P. (1977). *Understanding Rawls.* Princeton: Princeton University Press.