EXPRESSING OUR FALLIBILITY: A CONCEPTION OF PUBLIC REASON

Anthony David Taylor
Nuffield College
University of Oxford



Thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy in Politics in the Department of Politics and International Relations

Trinity Term 2017

70,152 words

Abstract

This thesis is about the reasonable agreement principle, a principle which holds that the exercise of political power must be acceptable to all reasonable citizens in order to be morally legitimate. Though this principle has become popular in contemporary political philosophy, it has been formulated and defended in a variety of often conflicting ways.

I argue first that a successful defence of the principle will have to meet three conditions. First, it must explain who reasonable citizens are. Second, it must offer a compelling a rationale for tying the legitimacy of the exercise of political power to what these citizens accept. Third, it must show that the rules or principles that would be acceptable to reasonable citizens are not implausible. In the first part of the thesis, I examine some of the most significant ways in which the principle has been formulated and defended, and argue that none meets these three conditions.

In the second part of the thesis, I develop an account of the reasonable agreement principle which can meet these three conditions. I argue that reasonable citizens should be understood as agents in circumstances where their powers of moral judgment operate free of distortions, offer an account of what these circumstances consist in, and suggest that a compelling rationale for the principle can be given when they are understood in this way. I then go on to consider what citizens in such circumstances would accept, arguing that they would accept principles of political morality that express a commitment to the fact that they are fallible choosers of their final ends.

Acknowledgements

I would first like to thank my supervisor, Simon Caney, for all of the feedback and advice he has provided over the last four years. Simon has been patient and encouraging while I have developed the ideas in this thesis, while also providing much sharp and useful criticism. I have benefitted greatly from both his guidance and from his extensive knowledge of political philosophy and related fields.

David Miller has also gone above and beyond his duties as a college advisor in providing support and direction at key points. I am grateful to him both for this, and for his role in making Nuffield College an ideal environment in which to write a thesis in political philosophy.

I also owe special thanks to Paul Billingham, Ian Carroll and Joe Horton. They have each provided generous amounts of written feedback, and I have learnt much from them in conversation.

In addition to the above, the following people gave helpful written feedback on one or more of the chapters of the thesis: David Birks, Dan Butt, Will Dahlgreen, John Halstead, Sam Kiss, Matt Kramer, Dan McDermott, Tom Parr, Jon Quong, Asbjørn Schmidt, Tom Sinclair, Zofia Stemplowska, and two anonymous reviewers for *Politics, Philosophy and Economics*.

I presented versions of the chapters of the thesis in the following places. The Braga meeting on Ethics and Political Philosophy at the University of Minho in 2016; the Hoover chair at the Université Catholique de Louvain in 2017; the Manchester Centre for Political Theory workshops in 2014, 2015 and 2016; the Nuffield Political Theory workshop in 2014 and 2016; and the Pavia Graduate Conference in Political Philosophy in 2015. I thank the audiences on all of these occasions for their questions and comments.

The first three years of work on this thesis were funded by a studentship from the Arts and Humanities Research Council, and the final year was funded by a scholarship from the Society for Applied Philosophy. Along the way I also received financial support from the University of

Oxford's Department of Politics and International Relations and from Nuffield College. I am grateful to all of these institutions, without which this work would not have been possible.

Finally, I thank my partner, Loren, and my parents, for all of their love and encouragement over the last four years.

Contents

Chapter 1: **Introduction** 9

PART I

Chapter 2: **The Reasonable Agreement Principle** 22

Chapter 3: **Reasonable Agreement and the Reactive Attitudes** 53

Chapter 4: **Rawlsian Political Liberalism** 78

PART II

Chapter 5: **The Publicity Requirement Defended** 114

Chapter 6: **Convergence, Fallibility, Satisfaction** 146

Chapter 7: **Conclusion** 185

Bibliography 193

1. INTRODUCTION

1.1. The Subject Matter of the Thesis

As it is commonplace to point out, the institutions and officials that make up contemporary states perform actions that would, if they were performed by a private individual, strike us as deeply problematic. They use force and the threat of force to ensure compliance with their rules, and they require those in their jurisdiction to provide material support for their activities. Now, a central question in political philosophy is: When is the exercise of political power legitimate? This question asks what the basis is for taking actions such as these to be morally permissible, when it seems that in many other circumstances they would not be so. This question is of interest not only because we want to know whether these practices of states can be vindicated, but also because, supposing that they can, understanding the scope of the legitimate exercise of political power is of great importance. What scope, if any, should be given to democratic procedures? What rights, if any, must be protected by a constitution? In order to answer such fundamental questions about the shape our political institutions should take, we need to have an account of what can render the exercise of political power legitimate in general.

This is the question I will discuss in this thesis. It is, as I will consider it, a question about political morality. It asks what the moral basis is for treating actions that would normally be taken to be morally impermissible to be permissible when performed by political institutions for certain reasons. And it asks what this moral basis can tell us about the proper scope of a legitimate state's powers. Though there are other senses of the term legitimacy that are used in discussions about politics, I will not be talking about them here.[1] The question is also about moral permissibility rather than other ways in which we might categorise exercises of political power from the point of

[1] The two most common alternative senses of the term are from Max Weber, who analyses legitimacy as a positive rather than a normative phenomenon, and understands it as consisting in the existence of various pro-attitudes towards political authority: "Politics as a Vocation," in *From Max Weber: Essays in Sociology*, (London: Routledge, 1991). And from Bernard Williams, who does analyse legitimacy as a normative phenomenon, but takes its normative force not to depend on a moral conception: *In the Beginning was the Deed: Realism and Moralism in Political Argument*, (Princeton, NJ: Princeton University Press, 2005).

view of a theory of political morality. We might also ask whether the exercise of political power is just or unjust, and perhaps also whether it is good or bad. I leave open, at this stage, what the relationship is between these various modes of evaluation. Perhaps a particular exercise could be legitimate while also being unjust. Perhaps legitimacy is solely a matter of following the right procedures, whereas justice and the good are substantive matters. I will not assume either of these views at the outset. Finally, as well as being distinct from questions about other ways of evaluating exercises of political power, the question is also distinct from that of what the basis is, if there is one, for a citizen's obligation to obey the law. I set this issue entirely aside.

Though there are a number of principles and ideals that might be thought to explain when and why the exercise of political power is legitimate, I will only discuss one here. I will refer to it as the reasonable agreement principle, though it has also been known as: the 'public justification principle', the 'public reason requirement', the 'reasonable acceptability requirement', or sometimes even just the 'liberal requirement'.[2] This principle holds that the exercise of political power is legitimate only if it would be acceptable to all reasonable citizens. In often quoted lines, the principle has been called both a foundational commitment and moral lodestar of liberalism.[3] However, as will begin to become clearer in chapter 2, since the reasonable citizens to whom this principle refers are an idealised constituency, versions of it differ radically both in how they specify the idealisation that is applied, and in their rationales for appealing to a principle of this sort in the first place. In this sense, then, any praise for or seeming widespread acceptance of the principle masks significant disagreement about its basis and specification.

It may seem odd that the thesis only discusses this one principle. What about other notions that might be thought to explain legitimacy? Must they not be discussed to, in order to answer our question? The primary reason for this focus is that I think that a version of the reasonable

[2] Though these terms for the principle are often used interchangeably, the term 'public reason' is also sometimes reserved for versions of the principle which require consensus on the reasons for political decisions rather than allowing for convergence on particular decisions for different reasons. I will introduce this distinction in section 2.2., but it can set aside for the time being.

[3] For the former, see Jeremy Waldron, "Theoretical Foundations of Liberalism," *The Philosophical Quarterly*, 37(147), 1987, 128. For the latter, see Stephen Macedo, *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism*, (New York, NY: Clarendon Press, 1991), 78.

agreement principle can explain the scope of the legitimate exercise of political power, but I do not think that the arguments needed to establish this conclusion are available in the literature. In order to show this, I will first need to clear the ground by examining the primary defences of the principle, which I do in part I. I then need to set out my own defence, which I do in part II. As I say, there are a number of different versions of the reasonable agreement principle, with quite different underlying rationales, so this task turns out to be more varied than might initially be supposed.

1.2. The Aim of the Thesis

My primary aim in defending the reasonable agreement principle is to provide guidance for forming judgments about the legitimate scope of the exercise of political power. I take it that, though we have some prior judgments about what an account of these limits will include, there are also many issues about which we will be unsure. If we could uncover the principle or principles that best explain our considered judgments about this question, then we could provide ourselves with guidance for the cases in which we are unsure.

Seeking the principles that best explain our considered moral judgments is a fundamental component of the method of reflective equilibrium.[4] This method holds that we are to work back and forth between our considered judgments and the principles that might explain these judgments, modifying both until we reach an equilibrium state. The idea of reflective equilibrium will crop up at various points later in the thesis, at which stage I will have cause to say much more about it. I will therefore refrain from discussing it here, beyond signalling a general commitment to the idea that this is the appropriate method for answering the question that I have set myself.

We can see the strength and importance of the strategy of attempting to uncover such underlying principles by considering the issue of liberal rights. Though many people now endorse rights to freedom of expression, association and conscience in some form, there is no widespread agreement on the question of why they warrant our endorsement. This absence of agreement may

[4] For a description of this method see John Rawls, *A Theory of Justice: Revised Edition*, (Cambridge, MA: The Belknap Press of Harvard University Press, 1999), 18-19.

initially appear to be unproblematic. After all, we might think, so long as we agree that these rights ought to be protected, it makes little difference if we disagree about our reasons for protecting them. But, in fact, it makes quite a lot of difference how we understand the basis of these rights. The reason for this is that any apparent agreement on these rights tends to mask disagreement about their scope. For example, the endorsement of freedom of religion is compatible with holding a wide range of views about the exact kinds of behaviour that this right protects. Similarly, support for freedom of expression is compatible with a variety of views on what kinds of speech, if any, can be permissibly regulated. We can readily repeat these examples for other liberties, such as freedom of movement and the freedom to produce and exchange goods. How should these questions about the scope of our liberties be settled? It seems to me that they must inevitably be settled by reference to our basis for endorsing them in the first place. That is, if we support a set of liberties on utilitarian grounds, we shall want to know what specification of them is most appropriate for the goal of maximising utility. And likewise, if we support them on the basis that the kind of autonomous choice they make possible is a fundamental component of the human good, we shall want to know what configurations of their scope are needed for such choice.

Analogously, with the legitimate exercise of political power, we will want to know what specification of these rights, if any, should constrain state action. And I hope that, by uncovering the principle or principles that offer the best explanation of our considered judgments about legitimacy, we can make headway on the question of how these and other ideas ought to be specified. This is not to suggest that once we have defended a foundation that we find satisfactory we will have a determinate answer to every possible question about the scope of the legitimate exercise of political power. It remains possible, and seems likely, that there will be a degree of indeterminacy. However, the range of possible answers to these questions will hopefully be narrowed considerably.

1.3. An Overview of the Arguments

Part I

I begin in chapter 1 by setting out the reasonable agreement principle in more detail, and arguing that a defence of it will need to answer three questions. First, it must tell us who reasonable citizens are (the idealisation question). Second, it must provide a rationale for taking the responses of reasonable citizens to be dispositive for the legitimacy of exercises of political power (the rationale question). And third, it must tell us what exercises of political power reasonable citizens could accept (the content question). I set these questions out in chapter 2, and the remainder of part I is organised around them. Over the course of chapters 2-4 I examine some of the most prominent defences of the reasonable agreement principle with respect to their prospects of providing satisfactory answers to these questions. Ultimately, I argue that none of the defences I look at provide satisfactory answers to them.

My consideration of defences of the reasonable agreement principle begins in chapter 2 with Charles Larmore's view. Larmore was one of the first[5] to attempt to base a view in political philosophy on the reasonable agreement principle as I understand it.[6] Not unlike others who have examined his view, I argue that it fails to provide a satisfactory account of why reasonable citizens would accept any exercises of political power at all, and avoids anarchistic conclusions only via an ad hoc fix. Further, I argue that the idea of respect for persons, which Larmore appeals to as the foundation of his view, has a significant case to answer in order to be capable of providing a basis for the reasonable agreement principle. After evaluating Larmore's view, chapter 2 also considers a more recent defence of the principle on the basis that it makes possible a valuable form of relationship: civic friendship, or political community.[7] Here I argue that while it is plausible to suppose that civic friendship requires citizens to share something in common, no plausible account

[5] I do not consider the defences of the principle in chronological order. I begin with Larmore's view as it is an effective way to frame the overall discussion, though I save Rawls's view (which was the first account of the principle as I understand it) for the end of part I, as it leads most effectively into the arguments of part II. In between, I discuss more recent formulations and defences of the principle.

[6] Principally in his *Patterns of Moral Complexity*, (New York, NY: Cambridge University Press, 1987).

[7] The most significant statement of this view to date is Andrew Lister's *Public Reason and Political Community*, (London: Bloomsbury, 2013), chapter 5.

of why the reasonable agreement principle is necessary to realise this relationship has been provided in the literature.

Chapter 3 moves on to consider one of the most well developed and discussed versions of the reasonable agreement principle, defended by Gerald Gaus.[8] Gaus's view is sensitive to the various issues that I raise for Larmore's view in chapter 2, and he explicitly aims to provide answers to each of the three questions that I argue a defence of the principle must answer. After setting out the view, I focus on its answer to the rationale question. Gaus's answer to this question is that we ought to accept his version of the reasonable agreement principle because it is presupposed by our moral reactive attitudes of resentment and indignation. He argues that these attitudes are so deep a part of us that we are unable to rationally reject them. And we are thus unable to rationally reject the reasonable agreement principle, which is presupposed by them. After setting out this argument, I examine the nature of our commitment to the moral reactive attitudes, and show that they need not be grounded in a commitment to the reasonable agreement principle. This shows, contra Gaus, that we can rationally reject that principle while maintaining a wholehearted commitment to our reactive attitudes. I conclude, then, that his view fails to provide a satisfactory answer to the rationale question.

The version of the reasonable agreement principle that I discuss the most in this thesis was developed by John Rawls, and I begin to discuss it in chapter 4.[9] Rawls's view has also been defended by a group of other philosophers who have introduced various supplementary and clarificatory arguments in support of it, and whose work I also consider in that chapter. After setting out the main the features of Rawls's view I again focus on his answer to the rationale question. A natural way of reading him points to his answer to this question appealing to an ideal of stability. On this account, the reasonable agreement principle is justified by its role in showing that a conception of justice could be stably complied in suitably ideal conditions without the need

[8] Most recently defended in *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, (New York, NY: Cambridge University Press, 2011).
[9] Principally in *Political Liberalism: Expanded Edition*, (New York, NY: Columbia University Press, 2005).

for excessive punishment.[10] I will call this the stability requirement, and it will be discussed throughout the thesis from chapter 4 onwards. Examining this requirement, I argue that, taken at face value, the relevant notion of stability is unfit to provide an answer to the rationale question. This is because there are a number of reasons to doubt that stability in this sense is a desirable property at all, or is sufficiently desirable to ground the reasonable agreement principle. On my understanding, the success of Rawls's argument for the principle will most likely depend on the defensibility of this stability requirement. And so I go on to consider what some of his defenders have said in support of this requirement, looking at arguments from Jonathan Quong, Larry Krasnoff and Paul Weithman.

Each of these arguments attempts to defend the stability requirement in a different way. Quong's argument ultimately appeals to the notion of probing the limits of practical possibility. On this view the aim of political philosophy is to set out an ideal of social cooperation that is realistically utopian: at the limits of what is practically possible.[11] A conception of justice must pass the stability requirement in order to confirm that it is its ideal achieves this aim. However, an appeal to the limits of practical possibility cannot provide the kind of defence of the stability requirement that would allow it provide support for the reasonable acceptability principle, or so I argue.

Krasnoff offers a more detailed argument in support of the stability requirement. He holds that it is a basic fact about normative theories in general that they aim at a kind of consensus: a consensus on whatever beliefs would produce the actions that they take to be right. He then argues that the pluralism faced by contemporary societies hinders this consensus, and since that pluralism is attributable to the full exercise of reason, our normative theory must be made consistent with it.[12] This argument has much to recommend it. However, I argue that Krasnoff's view is unable to explain which instances of pluralism are attributable to the full exercise of reason, and thus fails to

[10] The use of the qualifier excessive here will become clear in chapters 4 and 5, where I discuss this conception of stability in detail.
[11] *Liberalism without Perfection*, (New York, NY: Oxford University Press, 2011), 158-160.
[12] "Consensus, Stability, and Normativity," *The Journal of Philosophy*, 95(6), 1998, 269-292.

explain why we must satisfy the stability requirement as opposed to aiming for some other kind of consensus.

Weithman offers an appeal to reflective equilibrium, an idea that, as I mentioned above, will appear throughout this thesis. Weithman suggests that the state of affairs in which the stability requirement is satisfied is the state of affairs in which a general reflective equilibrium has been achieved, where 'general' denotes a state achieved by all agents together.[13] This merely pushes the question back a level, though, as what is now needed is a reason to think that reflective equilibrium must be general in this sense. I argue that there is nothing in the idea of reflective equilibrium itself or Weithman's argument that provides an answer to this question. I thus conclude that these arguments in support of Rawls's version of the reasonable agreement principle flounder on the absence of a satisfactory defence of the stability requirement.

My aim in part I of the thesis is not to provide a comprehensive survey of all of the defences of the reasonable agreement principle that have been elaborated and show that they all fail. Since there is now a large literature defending different versions of the principle, this would be a long and difficult task. My aim instead is to argue that a selection of the most discussed versions of the principle fail, in order to motivate the task that I begin in part II. And I take the work of Larmore, Gaus, and Rawls, alongside some of Rawls's key defenders to constitute a fair such selection. Having examined this selection in part I, I thus conclude that there is space for an alternative defence of the reasonable agreement principle to those currently on offer.

Part II

The first aim of part II is to provide a defence of the stability requirement. As I argue in part I, such a defence would allow us to provide an answer to the rationale question for a Rawlsian version of the reasonable agreement principle, which is the first and most fundamental question that a defence of that principle must answer. My argument in defence of the stability requirement in

[13] *Why Political Liberalism? On John Rawls's Political Turn*, (New York, NY: Oxford University Press, 2013), 335-339.

chapter 5 appeals centrally to a conception of the conditions in which our powers of moral judgment operate free of distortion. I argue that the conditions that a conception of justice must meet in order to satisfy the stability requirement are identical with the conditions in which our powers of moral judgment operate free of distortion. According to this argument, our conception of justice must satisfy the stability requirement in order to establish that we do not affirm it solely due to some distorting feature of our circumstances. The basic idea is this: we ought to evaluate our moral beliefs by considering whether we would continue to hold them in circumstances that abstract from certain features of our lives, such as our upbringing, which may have had a distorting influence on us, and we ought to affirm the conception that survives this abstraction. I take this to be a fairly typical component of moral reflection. A natural worry we might have about our moral beliefs is that they lack a rational basis because they are the product of various contingent features of our circumstances such as our upbringing. We sometimes reflect on the veracity our moral beliefs in light of this, asking ourselves questions such as: "Would I still think abortion were a grave moral wrong if not for my Catholic upbringing?"; or "Would I still think significant economic inequality was unproblematic if I hadn't grown up in this wealthy family?". I argue that, by setting out an appropriately idealised conception of an agent in particular background conditions, the stability requirement provides us with a point of view from which to evaluate what moral principles, if any, we would affirm in undistorted conditions. Chapter 5 defends this basic idea, and responds to a number of objections to it.

With this defence of the stability requirement to hand, the natural next question that arises is what principles, if any, could satisfy it. That there would be this kind of convergence even in ideal conditions can seem a miraculous possibility, and it is natural to wonder both whether it would occur, and whether any argument could show that it would. I begin chapter 6 by examining the most well worked out argument for the conclusion that a particular conception of justice satisfies the stability requirement: Rawls's Kantian congruence argument from *A Theory of Justice*.[14]

[14] See §86 for the argument itself, though it draws on all of part III of the book (347-514).

Though this argument is widely thought to be a failure, even by Rawls himself, I argue that by examining its general features and pinpointing the point at which it fails, we can begin to see what a successful argument for its conclusion would look like.

In order to construct an argument for the conclusion that a particular conception of justice satisfies the stability requirement, I begin with the idea of fallibility. That we are fallible choosers of our ends, and thus that our judgments about what ends are worth pursuing can be in principle mistaken, strikes me as a plausible candidate for view that could be the object of convergence among agents in idealised conditions with a diverse range of ends. Though each may disagree about which ends are worth pursuing, all can agree that they are fallible judges of this matter. In order to argue that we can get from this idea to the conclusion that a particular conception of justice satisfies the stability requirement I make use of two further ideas. The first is that from the acknowledgement that we are fallible choosers of our ends, we can infer that how advantaged we are might change, perhaps radically. We may come to occupy the position of a disadvantaged minority by changing our religious affiliation, to take one example. The second, roughly, is that there are conditions in which it is rational to give this possibility considerable weight in our practical reasoning. Drawing on these ideas, I argue that in order for a conception of justice to satisfy the stability requirement it will need to ensure that every social position is satisfactory from the point of view of its occupant.

The arguments of chapters 5 and 6 leave us with an account of the reasonable agreement principle with the following shape. In answer to the idealisation question, the reasonable citizens to which the principle refers are to be understood as agents in circumstances where their powers of moral judgment operate free of distortions. In answer to the rationale question, they are idealised in this way because we ought to accept the principles that we would affirm in if we were in such undistorted circumstances. In answer to the content question, the principles that would be accepted in such circumstances require us to ensure that the position of every member of society would be satisfactory from their own point of view.

I conclude in chapter 7 with a discussion of the status of ideals of the good and of justice under this version of reasonable agreement principle. With the arguments of chapters 5 and 6 to hand, I first ask whether can the principle can be affirmed alongside a full-blooded commitment to a particular conception of the good or broader worldview, and then, finally, I consider how it asks us to frame discussions about ideals of justice and political morality going forward.

PART I

2. THE REASONABLE AGREEMENT PRINCIPLE

2.1. Introduction

The first aim of this chapter is to introduce the reasonable agreement principle in more detail than I have so far. After doing so, I argue that a successful defence of it will need to provide answers to the three questions that I briefly introduced in chapter 1. The remainder of the chapter is dedicated to evaluating two attempts to defend the principle. First I consider Charles Larmore's view, one of the earliest account of the reasonable agreement principle, which attempts to base the principle on a requirement of respect for persons. I then move on to consider an alternative view, which attempts to ground the principle on an ideal of civic friendship or political community. I consider some variants of both of these views, and argue that none succeed.

2.2. The Principle: Three Questions

According to the most general characterisation of the reasonable agreement principle, it holds that in order to be legitimate political power must be exercised in a manner that all reasonable citizens can be expected to accept. What this claim amounts to will depend, it should be clear, on how the idea of a reasonable citizen is to be understood. For example, one way of specifying the principle might be to say that a citizen is only reasonable if they accept the moral truth. However, this would render the principle trivial. A utilitarian could agree that we ought to exercise political power only in a manner that all reasonable citizens could be expected to accept, because every citizen, if they were reasonable in this way, would be a utilitarian. And a proponent of any other moral view could also agree for exactly the same reason.[15] Therefore, if the principle is to have any substantive content reasonable citizens must not be understood simply as those who accept the moral truth.

[15] For a similar comment made by T.M. Scanlon with reference to his contractualism, see *What We Owe to Each Other*, (Cambridge, MA: The Belknap Press of Harvard University Press, 1998), 189.

Another way of specifying the principle would be to simply take reasonable citizens to be the actual citizens of our society. That is, we might say that the exercise of political power is legitimate only when it is exercised in a manner that all actual citizens could be expected to accept. However, the beliefs and commitments of actual citizens are so diverse that it is clear that there is nothing that they could all be expected to accept. Some actual citizens may well deny even analytic truths, after all. So on this rendering, the principle would amount to little more than a justification of anarchism. It would say that we should never exercise political power, because however we exercise it—and on whatever basis—we can expect that there will be some who would reject our decision.

I think we should also reject this anarchistic rendering of the principle. For one, if the idea of a reasonable citizen simply refers to the actual citizens of our society then the idea of the reasonable is redundant. More importantly, though, the view that there are never good grounds for the exercise of political power is intuitively implausible. Though there are many important disagreements about how and why political power can be exercised, most people would accept at the very minimum that the protection of some limited rights—such as to life and security—provide a sound basis for the exercise of such power.

In order for the principle to avoid both of these issues, what is needed is a specification of the beliefs and commitments of reasonable citizens such that they are neither univocally committed to the moral truth nor possessed of such diverse commitments that there is nothing on which they agree. Put another way, reasonable citizens will need be understood as an *idealised* constituency. They must be idealised such that they have some beliefs and commitments that go beyond the beliefs and commitments of ordinary citizens. Thinking about the principle in this way puts it, I think, in the company of dispositional or response-dependent theories. Views of this kind have been formulated about colour, value, reasons, and various other phenomena. They hold that whether or not something has a particular property depends on how suitably idealised agents would respond to it. So whether something is red, or of value, or a reason to ϕ, depends on whether it is

disposed to evoke a particular kind of response from a suitably idealised agent. For example, it might be said that something is of value if, and only if, we would be disposed to value it under conditions of full imaginative acquaintance.[16] Or, it might be said that something is a reason for Alf to ϕ if, and only if, were Alf fully rational, he would want to ϕ.[17] I will not discuss whether either of these examples of a response-dependent view is plausible, my point is simply that the reasonable agreement principle shares this formal structure. It tells us that an exercise of political power has the property of being legitimate if, and only if, suitably idealised citizens could be expected to accept it. But what might the 'suitably idealised' conditions be, in this case? This is the first question that I think a defence of the principle must answer, and I will call it the idealisation question.

How should this question be answered? I have argued so far that one way of idealising these citizens would make the principle trivial, and that not idealising them at all will leave it unable to justify anything. Between these two ways of idealising there are a host of options. Reasonable citizens could be characterised by various formal conditions, such as coherence and taking the most effective means to their ends. They could also be understood as having various moral commitments that fall short of accepting the moral truth—such as a commitment to comply with principles of justice provided others do likewise, or a shared aim of finding rules that no one can reasonably reject.[18] Claims about the beliefs and commitments of reasonable citizens are likely to be at least part of an answer to the idealisation question. Though in idealising we can also make claims about various other things, such as the information available to the citizens or the background conditions that they are in.

Given that there are a variety of options, it would be unacceptable to pick idealising conditions arbitrarily. A rationale will have to be given for the choice to idealise in one way rather than another. Whether a particular answer to the idealisation question is plausible will depend, to

[16] David Lewis, "Dispositional Theories of Value," *Proceedings of the Aristotelian Society*, Supplementary Volume 63, 1989, 113-137.
[17] Michael Smith, *The Moral Problem*, (Oxford: Blackwell Publishing, 1994), 130-180.
[18] Scanlon, *What We Owe to Each Other*.

a significant degree, on how compelling this rationale is. The second question that a defence of the principle must answer is thus: why are reasonable citizens idealised in this way? I will call this the rationale question.

It is sometimes suggested that the rationale for the reasonable agreement principle is simply that intelligent and well-meaning citizens tend to disagree about many matters: religion, the nature of the good life, justice, and so on. It is clearly true that intelligent and morally motivated people disagree in these ways, but it does not follow from this that we ought to only exercise political power in ways that intelligent and well-meaning citizens could be expected to accept. To move from the descriptive claim that persons with certain features tend to disagree to the normative conclusion that this constrains the legitimate exercise of political power will always require further argument. For this reason, however the idealisation question is answered, the rationale for the reasonable agreement principle can never be the phenomenon of reasonable disagreement itself. What an answer to the rationale question needs to do is to explain why it matters that persons with certain features tend to disagree. Simply restating that they disagree, or tend to disagree, can never provide a satisfactory answer to that question.

For this reason, I also think it is a mistake to focus on the ordinary language meaning of 'reasonable' when examining the principle. When applied to a person the term is sometimes used in ordinary language to mean that they are intelligent and well-meaning to a sufficient degree.[19] However, a principle that said that the exercise of political power must be acceptable to all intelligent and well-meaning citizens would be unappealing: much like actual citizens, there is likely nothing that intelligent well-meaning citizens could be expected to accept unanimously. Instead, I think the term should be understood as referring—as in a more archaic usage—to someone possessing the faculty of reason. Of course, this leaves it an open question quite what is involved in having this faculty in the sense that is relevant to the reasonable agreement principle. That is what an answer to the idealisation question must aim to tell us.

[19] The top entry for 'reasonable' in the *Oxford English Dictionary* is "within the limits of what it would be rational or sensible to expect; not extravagant or excessive; moderate," (Oxford: Oxford University Press, 2010).

Before introducing the third question that a defence of the reasonable agreement principle will need to answer, I want to make two brief comments. The first is that, though the reasonable agreement principle shares the formal structure of response-dependent views, it is perhaps more natural to see it as being a member of the family of contractarian or social contract theories. Those theories suggest that the content of morality or justice can be understood as what rational agents would accept under certain idealised conditions, and they have been defended by Rousseau, Kant, Rawls, David Gauthier, T.M. Scanlon, Jürgen Habermas, and others.[20] It is quite right to suggest that the reasonable agreement principle belongs firmly in this family. Indeed, these views will also have to answer the idealisation question and rationale question: they must give an account of the rationality of the agents and the conditions they are in, and explain why these particular norms and conditions were chosen. This raises a question about the scope this thesis. Why limit the discussion to the reasonable agreement principle, rather than considering all contractarian and social contract theories? A somewhat unsatisfactory answer to this question is simply that there is now a significant literature on the principle alone, and to consider every social contract theory would be too large a task. But more importantly, as I mentioned in the introduction, the primary aim of this discussion to motivate my own defence of the reasonable agreement principle, which will begin in the second part of the thesis. In order to provide this motivation, I think it is sufficient to show that some of the major recent ways of formulating and defending the principle are unsuccessful, so no exhaustive analysis of social contract theory is necessary.

The second note is on the way I have presented the principle and framed the key questions that need to be answered about it. I have focused on the question of how we idealise the constituency of reasonable citizens that the principle makes reference to, which I think is the most fundamental question that a defence of it must answer. However, others writing about the principle

[20] Jean-Jacques Rousseau, *The Social Contract and the Discourses*, (London: Everyman, 1993[1762]); Immanuel Kant, *Groundwork of the Metaphysics of Morals*, Edited by Mary Gregor and Jens Timmermann, (New York, NY: Cambridge University Press, 2012[1785]); John Rawls, *A Theory of Justice*; David Gauthier, *Morals by Agreement*, (Oxford: Oxford University Press, 1999); T.M. Scanlon, *What We Owe to Each Other*; Jürgen Habermas, "Discourse Ethics: Notes on a Program of Philosophical Justification," in *Moral Consciousness and Communicative Action*, (Cambridge, MA: MIT Press, 1990), 43-115

have sometimes suggested that different questions are fundamental. For example, Andrew Lister asks

> Is it *coercive state action* that must be unanimously acceptable to the appropriately idealized public; otherwise, we default to not acting, that is, not having a law or policy? Or, is it our principles, that is, our *reasons for political decisions* that must meet the unanimity criterion; otherwise, we default to not including the reasons in the grounds of our deliberations?[21]

He points out that it makes a significant difference to what the principle recommends whether we understand it as a constraint on coercion, or as a constraint on reasons for political decisions.

In a similar vein, others have focused attention on the question of whether the principle requires convergence or consensus.[22] That is: is the principle satisfied if reasonable citizens converge on accepting a particular law or policy for different reasons? Or must there also be consensus on the supporting reasons themselves?

For my purposes both of these questions are premature. If there is no plausible answer to the idealisation question and the rationale question, then the principle is indefensible, regardless of whether it is best understood as a constraint on coercion (or reasons) or as requiring consensus (or convergence). Moreover, it seems to me that if we *do* have acceptable answers to the idealisation question and the rationale question, then these will provide us with some guidance on how we ought to specify the principle.

Let me now say something about what we might be called the content of the principle. That is: what kind of claims could satisfy it? A full defence of the principle will have to provide an answer to this question, which I will call the *content question*. It will have to tell us what rights, principles of justice, political procedures and other norms would be acceptable to all reasonable citizens. I have already noted one way in which the content of the principle acts as a constraint—namely, that some ways of answering the idealisation question would render it no more than a justification for anarchism. However, it would not be sensible, I think, to say in advance of answering the idealisation question and the rationale question that the content of the principle

[21] Andrew Lister, *Public Reason and Political Community*, (London: Bloomsbury, 2013), 1. Emphasis in original.
[22] This distinction is foregrounded in the work of Gerald Gaus and Kevin Vallier. For an instructive account, see Vallier's "Convergence and Consensus in Public Reason," *Public Affairs Quarterly*, 25(4), 2011, 261-279.

generates many other constraints. Though the principle will no doubt be more palatable to some person if it turns out to provide support for his or her prior beliefs, my aim in defending it is in part the hope that it might be able to provide guidance in cases where we are uncertain. Given this aim, I will be sparing in making claims about what the content of the principle must accommodate in advance.

2.3. Larmore on Neutrality

Charles Larmore was one of the first to embark on the elaboration of the view known as political liberalism, which puts the reasonable agreement principle at its centre. His account has, I believe, had a significant impact on the lay-philosophical understanding of the principle in the literature: I suspect that those who approach the reasonable agreement principle for the first time often suppose that it has something like the shape and basis that Larmore gives to it. For this reason, I think there is some value in considering his view first. Identifying some of its weaknesses can serve to illustrate some common pitfalls faced by elaborations of the reasonable agreement principle.

Before I delve into Larmore's view I wish to make a preliminary comment. The philosophical debate about the reasonable agreement principle and its defensibility is often embedded in a wider debate between perfectionists and anti-perfectionists or liberal neutralists. The anti-perfectionist or neutralist view is characterised by the claim that conceptions of the good should not play any central role in political justification. A conception of the good is an account of what is of value or worthwhile or the object of rational desire or a contributor to flourishing. Anti-perfectionists thus hold that state action should not be based on judgments about which of these conceptions is worthy of pursuit. Instead, it should aim at bringing about a conception of justice that allows individual citizens to pursue their own conception of the good under fair conditions.[23] By contrast, their opponents deny this restriction, allowing for state action to be based on claims about the veracity of conceptions of the good in various ways. Though I have not foregrounded

[23] For an instructive account of the distinction between the two views, see Jonathan Quong, *Liberalism with Perfection*, 2-4.

this debate in introducing the arguments of the thesis so far, the reasonable agreement principle often plays a central role within it. Wielding the principle, the anti-perfectionist side in this debate have argued that reasonable people disagree about which conception of the good is the right one, and therefore that the state ought to eschew appeals to conceptions of the good in political justification thus remaining 'neutral' on their veracity. I introduce this debate here as it is necessary to understand some of the discussions about the reasonable agreement principle and the language that those discussions are couched in. However, as should be clear from my discussion so far, whether or not the principle can be wielded by the anti-perfectionist side in this way will depend on the specifics. As critics of anti-perfectionism have been keen to point out, it is far from obvious that reasonable citizens agree about justice but disagree about the good.[24] Whether or not the distinction between what is and is not acceptable to reasonable citizens tracks the distinction between conceptions of the good and principles of justice depends on how the idealisation, rationale and content questions can be answered.

Moving on to Larmore's view, I will begin with his answer to the idealisation question, which is relatively simple. He takes reasonable persons to be those "thinking and conversing in good faith and applying, as best [they] can, the general capacities of reason which belong to every domain of enquiry".[25] Since over the course of the thesis I will talk about various ways of specifying reasonable persons, I will refer to those idealised in this way as L-reasonable. Larmore's principal claim about L-reasonable persons is that they will not accept a single conception of the good. This claim needs little by the way of argumentative support: it is relatively clear, after all, that those exercising their capacities as best they can will not come to unanimous agreement on a single conception of the good.

[24] This thought appears as the so-called asymmetry objection, proponents of which hold that anti-perfectionists lack a good basis for their asymmetrical treatment of justice and the good, as well as more specifically in the claims of those who argue that some uncontroversial claims about the good life would be the object of agreement among reasonable citizens. For the former, see: Simon Caney, "Anti-Perfectionism and Rawlsian Liberalism," *Political Studies*, 43(2), 1995, 248-264; Michael Sandel, *Liberalism and the Limits of Justice: Second Edition*, (Cambridge, Cambridge University Press, 1998), 202-210; Jeremy Waldron, *Law and Disagreement* (Oxford: Oxford University Press, 1999), chapter 7; among others. For the latter, see: Andrew Lister, *Public Reason and Political Community*, 44-53 and Franz Fan-lun Mang, "Liberal Neutrality and Moderate Perfectionism," *Res Publica*, 19(4), 2013.

[25] Charles Larmore, "Political Liberalism," *Political Theory*, 18(3), 1990, 340.

Since L-reasonable persons disagree about the good, and Larmore accepts the reasonable agreement principle, he thinks that decisions about how political power is to be exercised ought to be neutral with respect to conceptions of the good. That is, the justification for a law or policy should not be based on a conception of the good about which reasonable persons disagree. Following Larmore, I will call this the Principle of Neutrality.[26] Larmore further holds that the basis of the Principle of Neutrality must itself be neutral, or as neutral as possible.[27] The idea here is that we do not simply need to be neutral between the different L-reasonable conceptions of the good that citizens affirm, but that our reason adopting a policy of neutrality must itself be one that L-reasonable citizens could be expected to accept.

What possible basis for the Principle of Neutrality could satisfy this constraint? Larmore answers this question by appeal to the following two norms.

> Norm of Rational Dialogue: When we encounter points of disagreement with others, we should respond by retreating to shared beliefs to solve our dispute.
>
> Norm of Equal Respect*:* We must not bring about compliance with our principles solely by force or the threat of force.[28]

The rationale for the first norm is pragmatic. If you want to address those who disagree with you, and reach a solution to your disagreement via rational discussion, you will need to appeal to some common ground. If you have the goal of settling a disagreement via deliberation, then it will be ineffective to cling to premises that your interlocutors reject. However, as Larmore notes, this presupposes that you *want* to settle your disagreement via rational discussion.[29] Why should you aim to settle your disagreements in this way rather than, say, by force? Norm (2) answers this question. Settling your disagreements by forcing others to comply with your terms would be failing

[26] Ibid., 342.
[27] Ibid.
[28] Ibid*.,* 347.
[29] Ibid., 348.

to treat them with equal respect, by bypassing rather than engaging their distinctive reasoning capacities.[30] So in order to treat others with the respect they are owed, you must settle your disagreements with them via rational dialogue rather than coercion, and the most effective way to do this is by retreating to shared beliefs when you encounter disagreement. This is a neutral defence of the principle, according to Larmore, because its compatible with a range of the conceptions of the good held by L-reasonable persons.[31]

Returning to Larmore's answer to the idealisation question, an important point to note about this answer is that whether or not a view is L-reasonable depends entirely on the process by which it was arrived at.[32] Provided that I affirm the view I do as an outcome of applying, in good faith, my capacity to reason about the subject matter, that affirmation is reasonable. Understanding the reasonable in this way means that in principle any view could be reasonable. If Alf exercises his powers of judgment as best he can, and reaches the conclusion that he is morally permitted or even required to kill those who do not adhere to his conception of the good, then—provided he has genuinely satisfied the condition of reasoning in good faith about the issue—this account is forced to say that his view is reasonable.[33] Setting aside the thorny issue of how one might work out whether or not others are genuinely exercising their powers of judgment in good faith, the most natural worry to have about a view that understands the reasonable in this way is this. Why should we think that there will be anything at all that L-reasonable citizens would unanimously accept? After all, as the example of Alf shows, we cannot guarantee that citizens idealised in this way will accept even the most basic of moral norms against killing and harming, let alone anything like a conception of basic rights. If that is right, Larmore's view will be no more than a justification for anarchism, and so will violate the minimal condition that I placed on answers to the content question above.

---

[30] Ibid., 347.
[31] Ibid., 346.
[32] For this way of characterising Larmore's view, see Gerald Gaus, "The Rational, the Reasonable and Justification," *The Journal of Political Philosophy*, 3(3), 1995, 245.
[33] An example of this form is used by Jonathan Quong against Gaus's conception of the reasonable, which is also procedural. See his "What is the point of public reason?" *Philosophical Studies*, 170(3), 2014, 547.

I will now consider how Larmore attempts to respond to this worry, and in the process I will draw out some of the other distinctive features of his view. Stated fully, my concern here is that when the Norm of Rationale Dialogue is applied and a retreat to shared beliefs is attempted, in some cases there will be no common ground. This is because, as I have pointed out, L-reasonable citizens could in principle hold any set of beliefs, so there is no guarantee that there will be common ground beyond the bare commitment to working out our views as a good faith exercise of judgment. If there is no common ground between reasonable citizens, there will be no way to bring about compliance with any principles that is consistent with the Norm of Equal Respect. In such a case, there will be no principles that can govern cooperation with our co-citizens that show appropriate respect for them.

Aware of this difficulty, Larmore suggests that we should trade-off the value of satisfying the Principle of Neutrality against other values. We do this by introducing, where there is no common ground, just as much material as is necessary to provide a basis for choosing principles in a manner consistent with equal respect. The underlying idea here is that "one should institute only the least abridgement of neutrality necessary for making a decision possible".[34] In admitting material in order to make such a trade-off, he claims we must satisfy the following two rules.

> Rule 1: Admit beliefs that are the least central to any L-reasonable conception of the good life.
>
> Rule 2: Admit beliefs that the least number of L-reasonable persons do not hold.[35]

These rules describe two exceptions to the Norm of Rational Dialogue. The first rule states that where the parties share no common ground they should refrain from removing from consideration beliefs that, though still the object of disagreement, are the least central to any of their conceptions of the good. This implies the following: in cases where a policy is supported by one group on the

[34] Charles Larmore, *Patterns of Moral Complexity*, 67-8.
[35] Ibid. 68.

basis of a belief that is a central component of their conception of the good, but opposed by another group on the basis of a belief that is only a peripheral component of their conception of the good (and assuming, for the sake of argument, that everyone falls into one of these two groups), the policy in question ought to be enacted.[36] On this view, citizens' support for or opposition to a policy is weighted according to how central it is to their conception of the good. If the L-reasonable opposition to a policy of some citizens is based on peripheral beliefs, this opposition can in some cases be permissibly set aside. This is in keeping with the more general thought that we should restrict the principle of neutrality as little as we possibly can, while making some decision possible. Our underlying rationale for being neutral between L-reasonable conceptions of the good is the Norm of Equal Respect, and it is a less egregious violation of this norm to set aside a peripheral belief than a central one. Larmore takes appealing to Rule 1 to be preferable, but thinks that it will often prove impossible to implement.[37] In these cases, he thinks we should appeal to Rule 2, which tells us to minimise the number of people who have a complaint against the law or policy based on an L-reasonable conception of the good. Again, the underlying idea here is to minimise the violations of the principle of neutrality, if not by severity, then by number.

Let me return now to the concern about Larmore's answer to the content question. That concern was that because L-reasonable citizens could in principle have any beliefs, there might be no beliefs that they all share. In such a case, there will be no principles with which we can bring about compliance consistent with the Norm of Equal Respect. The preceding discussion shows how Larmore's view attempts to address this concern. By countenancing some limited violations of the principle of neutrality in the line with the two rules above, he can argue that there will always some principles on which it is permissible to bring about compliance.

However, I think that this move is objectionably ad hoc. If a commitment to the Norm of Equal Respect leads to the view that there are no principles with which it is permissible bring about

[36] James S. Fishkin, "Review: *Patterns of Moral Complexity*," *Political Theory*, 17(1), 1989, 153-156; and Jonathan Seglow, "Neutrality and Equal Respect: On Charles Larmore's Theory of Political Liberalism," *The Journal of Value Inquiry*, 37(1), 2003, 83-96.
[37] *Patterns of Moral Complexity*, 68.

compliance, because there is too little on which L-reasonable people agree, why should violations of the norm be permitted in order to avoid this conclusion? The rationale that Larmore gives here is that the value of satisfying the Principle of Neutrality must be weighed against the value of some decision being made about the principles governing basic liberties and the distribution of goods. However, the conclusion that there ought to be no rules or principles in a particular domain *is* a decision being made about the principles governing basic liberties and the distribution of goods—it is simply the decision that there should not be any rules or principles because there is no appropriately neutral basis for them. It is therefore not clear if there is any sense in which this view amounts to weighing the value of a decision being made against the Principle of Neutrality in countenancing violations of the principle in accordance with the two rules. If Larmore is not weighing this other value against the value of neutrality, then he is introducing an ad hoc exception to the principle to avoid its potentially anarchist implications.

The only way I can see to avoid this objection is by saying that what is being weighed against the Principle of Neutrality is not the value of a decision being made, but rather the value of protecting some basic liberties or having an appropriate distribution of goods. That is, the value of bringing about a particular conception of justice or morality. This would explain why the decision to have no norms or principles in this domain would be unacceptable: it is unacceptable because there are some norms or principles that ought to be enforced. However, this move would leave the Principle of Neutrality and the Norm of Equal Respect unable to explain the scope of the legitimate exercise of political power. No question about whether some particular exercise of political power would be legitimate could be settled solely by appealing to the Principle of Neutrality and the Norm of Equal Respect, because in order to answer such a question we would need to examine the value that is being weighed against the former principle. Given the goal of explaining when the exercise of political power is legitimate, this is a devastating cost.

2.4. Larmore on Treating Others as Mere Means

This objection suffices to show that Larmore's view is inadequate for my purposes. However, in pressing it I have not yet engaged with his claim that the basis for the reasonable agreement principle is equal respect. What are the prospects for an appeal to equal respect doing this job? I think there is a severe initial difficulty with this possibility. The thought that we ought to show equal concern or respect for others is appealed to fairly often in moral and political philosophy. However, a natural understanding of the idea is that you treat people with such respect when you act in accordance with appropriate principles of morality or justice. This point is put well by Rawls when he claims that "without the principles of right and justice, the aims of benevolence and the requirements of respect are both undefined; they presuppose these principles already independently derived".[38] If respect is understood in this way it will be unable to provide an answer to the rationale question. If treating people with respect is just treating them in accordance with the correct principles, then the idea of respect will be unable to explain what the correct principles are. It simply assumes these are already to hand. When examining any appeal to equal respect this thought must be kept in mind, as clearly the notion will have to be formulated such that it has some content independent of principles of justice in order to be able to serve this purpose.

Larmore introduces the Norm of Equal Respect as being 'closely related' to Kantian claim that we should never treat persons merely as a means to our ends, but rather always as ends in themselves.[39] For Kant, this claim was one of his three formulations of the categorical imperative—an imperative that applies to all rational agents regardless of what particular ends they have. What avoiding treating others merely as a means to your ends involves is a matter of dispute. We often treat others as means to our ends in ways that are clearly innocuous, such as when we ask them for assistance and they agree to help us. However, what is wrong according to the Kantian claim is treating others *merely* as a means to our ends. On one understanding, we treat others merely as a means when we use them as a means to our ends while giving insufficient or no weight to their

[38] Rawls, *A Theory of Justice*, 513.
[39] "Political Liberalism," 348.

moral claims or wellbeing.[40] Few would deny that we ought to avoid treating others in this way. But does this imply that the justification for legislation should not be based on claims about which reasonable persons disagree? The idea seems to be—as Thomas Nagel has put it—that "if you force someone to serve an end that he cannot share, you are treating him as a mere means".[41]

In order for it to be the case that forcing someone to serve an end that they reasonably reject necessarily gives too little weight to their moral claims, they must have a moral claim against being treated in this way. If they do not have such a claim, then we could force them to serve an end that they reasonably reject while giving sufficient weight to all of their valid moral claims. This means the question becomes: do persons have a claim against being forced to serve an end that they can reasonably reject? Perhaps they do. But the appeal to equal respect was meant to explain *why* we ought not coerce persons on the basis of claims that they can reasonably reject. If equal respect is to be understood as closely related to the Kantian claim that you ought never treat another merely as a means to your ends, and what this amounts to is avoiding treating others in ways that give insufficient weight to their moral claims, then there will be no non-question begging way for it to do this explanatory work. The principle that we should not treat others as mere means to our ends does not establish that persons have claims of this sort.

Maybe there is an alternative formulation of the idea that could do this explanatory work, though it is not clear what it would look like. What makes this difficulty seem insurmountable for Larmore is that his rationale for appealing to this conception of equal respect is that it is a defence of the Principle of Neutrality that all or most L-reasonable persons would accept. But the reason the idea of equal respect appears to be one that most L-reasonable persons would accept is that it does not specify what failing to give sufficient weight to the moral claims of others amounts to, because it does not tell us what moral claims persons have. In order for an alternative formulation of the idea of equal respect to provide a non-question begging explanation of why we ought to

[40] Derek Parfit, *On What Matters: Volume One*, (New York, NY: Oxford University Press, 2011), 212.
[41] "Moral Conflict and Political Legitimacy," *Philosophy & Public Affairs*, 16(3) 1987, 223.

accept the Principle of Neutrality, it would have to remove this lack of specificity. By doing that it will inevitably lose its status as something most L-reasonable citizens would accept.

2.5. Barry on Respect and Scepticism

Brian Barry also expressed scepticism that Larmore's formulation of the Norm of Equal Respect can do the work required of it, though he suggested that the view could be rescued by appealing to a particular kind of scepticism or uncertainty about conceptions of the good.[42] He begins by asking why I cannot satify the Norm of Equal Respect by simply explaining to those I coerce why my conception of the good is true. After all, I might think that my conception of the good is one that every reasonable person should be prepared to accept. He argues that this would be inadequate because "no conception of the good can justifiably be held with a degree of certainty that warrants its imposition on those who reject it".[43] His thought is that it is clear why equal respect requires neutrality among conceptions of the good once it is recognised that conceptions of the good are not the kind of thing in which we have sufficient confidence to impose them on others who reject them.

Are there good grounds to share Barry's scepticism or uncertainty about conceptions of the good? He makes the case for it by generalising from a discussion of religious toleration. Religious toleration is often understood to have risen out of the intractability of religious disagreements in Europe after the Reformation, as it became clear that there was little chance of converting everyone to same view. The support this provides for scepticism is that given "so many people have devoted so much effort over so many centuries […] with so little success in the way of securing rational conviction among those not initially predisposed in favor of their conclusions," it is hard to believe there could be a single religious conception that all reasonable persons ought

[42] Brian Barry, *A Treatise on Social Justice Volume II: Justice as Impartiality*, (Oxford: Clarendon Press, 1995), 164-177.
[43] Ibid., 169

to accept.[44] If this is true for religious conceptions, it is also true *a fortiori* for all conceptions of the good.

The problem with this argument is that, whatever its force, it gives us no reason to be sceptical about conceptions of the good in particular.[45] Why not also think that there is no conceptions of justice or morality that would be the object of rational or reasonable convergence? After all, conceptions of justice have also been the subject of intense debate and none have secured convergence. If we have no special reason to be sceptical about conceptions of the good in particular, then Barry's proposed modification would also imply that no conception of justice can be held with sufficient certainty to warrant imposing it on those who reject it. This means that taking this scepticism seriously implies that we can never legitimately exercise political power. Therefore, the addition of Barry's sceptical claim does nothing to help Larmore's view.

It is worth stating what I think are two lessons that come out of this discussion of Larmore's view. The first stems from the difficulty that his answer to the idealisation question generates for him. By holding that reasonable citizens are those "thinking and conversing in good faith and applying, as best [they] can, the general capacities of reason which belong to every domain of enquiry" [46], he gives an answer to this question that holds that in principle any view can be reasonable, provided it is the outcome of such a process. This feature of the view makes it difficult for a satisfactory answer to the content question to be given. If any view can be reasonable provided it the outcome of this process, then, as I have argued, it is likely that there is little or nothing that reasonable citizens unanimously accept, and therefore the reasonable agreement principle will justify few if any exercises of political power.

The second lesson is that it is difficult to appeal to the idea of respect to answer the rationale question. Such an appeal would have to find a way of formulating this notion that does not presuppose a prior account of what our moral claims are. None of this is to say that developing a

---

[44] Ibid. 171.
[45] Simon Caney, "Impartiality and Liberal Neutrality," *Utilitas*, 8(3), 1996, 277-278.
[46] "Political Liberalism," *Political Theory*, 340.

version of the reasonable agreement principle that uses a similarly idealised conception of the reasonable citizen or appeals to respect as a rationale is an impossible task. I doubt that there is any way of establishing such a general conclusion. But I do think that flagging up these difficulties and showing that the most prominent account with these features fails suffices as motivation to begin to look elsewhere.

2.6. The Civic Friendship Rationale

Setting Larmore's view aside, an alternative popular answer to the rationale question appeals to the value of civic friendship. Civic friendship is a particular relationship that, its defenders argue, can hold between co-citizens and shares some of the valuable features of friendship in general. There are a number of appeals to the value of civic friendship in the literature on the reasonable agreement principle. Though they differ in their precise details, the basic idea is fairly simple. There is a certain kind of political community or friendship that is prima facie desirable, but on reflection turns out to be impossible to achieve. This is a political community that is based on all citizens sharing a particular worldview or conception of the good life. Such a community is prima facie desirable because the relationships between citizens it makes possible share what are arguably some of the important features of friendship: shared ends and projects, and a commitment to promoting one another's good understood in terms of the values that we share. However, in spite of this prima facie desirability, this ideal of community or friendship cannot be achieved in a political society. This is because, familiarly, the reasonable citizens of liberal democratic societies hold a variety of different conceptions of the good. Given their inevitable disagreement about which conception of the good is the right one, the ideal of a community based on a single conception of the good is out of reach for such citizens.[47]

Fortunately, all is not lost according to the defenders of the civic friendship rationale. If the reasonable agreement principle is followed, then we can have a different kind of political

[47] Kyla Ebbels-Duggan, "The Beginning of Community: Politics in the Face of Disagreement," *The Philosophical Quarterly*, 60(238), 2010, 52.

community. And it can be one that retains some of the important features of a community based on a single conception of the good. By basing political justification on what reasonable citizens share, whatever that happens to be, we can retain the idea of shared ends and projects and perhaps a shared commitment to furthering each other's good, understood in a rehabilitated way.

In order to begin to make sense of this argument, I will first need to say more than I have said so far about the requirements that some defenders of the reasonable agreement principle take it to place on individual citizens and government officials, as these play a fundamental role in the mechanism by which the principle is meant to realise civic friendship.

2.7. Duties on Individuals and Government Officials

The thought that acting on the reasonable agreement principle allows us to realise a political relationship of civic friendship with our co-citizens was first suggested by Rawls:

> To make more explicit the role of the [reasonable agreement principle] as expressed in public reason, note that its role is to specify the nature of the political relation in a constitutional democratic regime as one of civic friendship. For this [principle], when government officers act from it in their public reasoning and other citizens support it, shapes the form of their fundamental institutions.[48]

In spite of this claim, it not obvious that Rawls intends for the idea of civic friendship to play the role of answering what I have called the rationale question. Since I will be discussing Rawls's view in detail in chapter 4, in this section I will primarily consider the civic friendship rationale in isolation from it. One exception to this isolation is that to understand how defenders of the civic friendship rationale take the reasonable agreement principle to realise this relationship, we need to have before us an understanding of the duties that some defenders of the principle, primarily Rawls, take it to impose on individual citizens and government officials.

For a number of defenders of the reasonable agreement principle it does not constitute what is sometimes called a criterion of rightness. That is, it does not simply say that right action in the domain of political morality is that which would satisfy the principle. Instead, they hold that it

[48] "The Idea of Public Reason Revisited," *Collected Papers*, (Cambridge, MA: Harvard University Press, 1999), 579.

imposes individual duties on individuals and government officials that are distinct from maximising the extent to which the principle is fulfilled. These are the duties to which Rawls is referring the quoted passage: citizens and government officials are to conduct their political deliberation and voting practices in terms that all reasonable citizens could be expected to accept. Put differently, they ought to comply with the *duty of civility* which directs them to "be able to explain to one another on those fundamental questions how the principles and policies that they advocate for can be supported by the political values of public reason".[49]

As we shall see, compliance with this duty is a fundamental part of proponents of the civic friendship rationale's account of how the reasonable agreement principle allows us to realise a valuable kind of relationship in the face of disagreement. I now turn to that account.

2.8. Reasonable Disagreement and Political Community

A.

Proponents of the civic friendship rationale begin with an account of the features of ordinary *non*-civic friendship. According to Kyla Ebbels-Duggan, friendship involves: acting co-operatively as equals; seeing each other as having the authority make claims us; not trying to manipulate or control each other; and not seeing each other "merely as obstacles to or tools for attaining our own ends".[50] Building on this, to be *civic* friends, we must act co-operatively with our co-citizens. This means reasoning together with them, and not seeing them as enemy combatants with whom we battle to impose our view.

Ebbels-Duggan argues that a commitment to the reasonable agreement principle and the duties for individual citizens and government officials outlined above follows from this ideal of friendship. If state action is based on considerations that some of our reasonable co-citizens reject, then our relationship will cease to be based on acting co-operatively with them as equals. Given that reasonable citizens disagree about the good life, this means that an ideal of political community

[49] Rawls, *Political Liberalism*, 217.
[50] "The Beginning of Community,", 56.

based on everyone sharing the same conception of the good is off the table, and that our alternative ideal must instead be based on whatever it is reasonable citizens hold in common. To reject civic friendship entirely in favour of the enemies-to-be-defeated view of our co-citizens would be unjust, as it fails to acknowledge that our co-citizens are a potentially authoritative source of claims.[51]

But as Ebbels-Duggan accepts, there are clearly a range of relationships that fall between that of co-operation based on the reasonable agreement principle and that of enemies-to-be-defeated. The latter does not take other citizens to be the potential source of moral claims at all, and there are plenty of ways in which we might take account of our co-citizens as potential sources of such claims that do not involve acting on the reasonable agreement principle or the corresponding duties I outlined. For example, we might see the interests of our co-citizens as something we must take account of, as giving us perfectly valid moral reasons, without taking this to commit us to justifying our political decisions by appeal to reasons that all reasonable citizens could be expected to accept. In spite of the fact that there are a variety of relationships that are not based on the reasonable agreement principle but also do not amount to seeing our co-citizens as enemies-to-be-defeated, Ebbels-Duggan takes co-operation on the basis of the reasonable agreement principle to constitute the appropriate ideal for civic relations.[52]

If this argument is best that can be said for the civic friendship answer to the rationale, then it fails: it depends on us facing a dichotomous choice between disregarding the interests of our co-citizens entirely and accepting the reasonable agreement principle, a dichotomy that Ebbels-Duggan confesses is false. Once we accept that there are range of ways to conduct civic relations that do take consideration of the claims of our co-citizens, we will want to know why the reasonable agreement principle even constitutes the ideal at all. The only conclusion that is warranted by this argument is that the reasonable agreement principle is one way of taking account of the moral claims of our co-citizens. And this conclusion is far too weak to support the reasonable agreement principle or its related duties.

[51] Ibid.
[52] Ibid., 57.

B.

Andrew Lister offers a more detailed account of civic friendship and its relationship to the reasonable agreement principle.[53] Acting on the duty of civility involves excluding some reasons from your deliberation about how to act in the political realm because they are controversial among reasonable citizens. Such an exclusion of reasons would also be justified, he argues, in an interpersonal relationship such as a marriage. In a marriage we may exclude from our deliberation reasons that our partner does not accept on the grounds that this allows us to act together and constitute a community in spite of our disagreements. Lister introduces the following case in order to develop this claim.

> *Private Education*: Colin and Jane disagree about whether to send their son to private school because they have different views about social justice and parental responsibilities. Colin believes strongly in public education, and thinks it is an illegitimate form of parental partiality to send one's child to private school when there is a decent public alternative. Jane does not share Colin's demanding conception of personal responsibility for social justice; she thinks they should enrol their son in the private school if doing so would be to his benefit. She also thinks that they should consider the benefits that their son would get simply from joining the club of elite school families, such as establishing a network of contacts. Colin, however, thinks these benefits are irrelevant to the decision.[54]

Given their disagreement, Colin and Jane might agree to bracket the reasons on which they are divided. Since Jane does not think that Colin's social-justice-based considerations are relevant to the question of where their son goes to school, those will be taken out of the picture. Likewise, since Colin does not think that Jane's elite-networking-based considerations are relevant, these will be bracketed too. Having set aside these contested reasons, they might now base their decision instead on what they agree is relevant to the decision, in this case, perhaps, just the intrinsic merits of the education itself. In this case, their commitment to act on only shared considerations can be "constitutive of a valuable relationship, in [the] face of the inevitable disagreement that crops [up] in the course of a common life".[55]

---

[53] *Public Reason and Political Community*, 105-133.
[54] Ibid., 107-108.
[55] Ibid., 108.

Lister's aim is to show that the reasonable agreement principle can be understood as a condition of political community in precisely the same fashion. He offers two principal reasons for holding that reason-bracketing would constitute a valuable form of community that he takes to apply to the relationships of co-citizens. The first is that common knowledge of the fact that public decisions are aiming to satisfy the reasonable agreement principle would be a valuable public good. Public goods are goods that are non-excludable and non-rivalrous: they cannot be provided to some citizens with being provided to all, and their use by some does not decrease their availability to others. The public good that common knowledge of a society aiming to satisfy the reasonable agreement principle constitutes is that no citizen "need think that public decisions aim at fundamentally alien purposes; everyone recognizes that others are committed to making decisions based on reasonably shareable grounds".[56]

The second reason is that Lister specifies is that bracketing controversial considerations and reasoning on the basis of what we each reasonably accept is a way of constituting ourselves as a collective agent.[57] Unlike the first reason, Lister thinks that this one has the virtue of explaining why it is essential that we only enact policies that are supported by reasons that are acceptable to all, rather than simply going for policies that are acceptable to all, even if there is no agreement on the underlying reasons.[58] There are a number of conceptions of group agency, but Lister takes it to require at least shared intentions and a level of collective rationality. Shared intentions are present when each member of a group intends that they, together, φ, as opposed to having intentions that merely overlap in their content. A minimum level of collective rationality is present when the behaviour of the group can be described as reliably promoting a purpose or set of purposes.[59] Beyond these conditions for group agency, Lister tells us that to constitute a political community

[56] Ibid., 110.
[57] Ibid., 113.
[58] This is significant for Lister in part because he aims to defend a version of the reasonable agreement principle that requires shared reasons in this way. See Ibid., 2.
[59] Ibid.

a group must also be such that each of its reasonable members can see it as acting for common goals, or at least attempting to.[60]

The significance of satisfying these conditions for collective agency stems again from the effect that common knowledge of decisions being made in this way has on the character of citizens' relationships with one another. It is at this stage that the fact the reasonable agreement principle requires agreement on the reasons for policy decisions rather than agreement on the policies themselves becomes important. In interpersonal relationships, we care not just about how people act, but why they act the way they do. This is because the reasons for which people act matter for what they would do in various counterfactual situations. Lister writes:

> I have a different relationship with someone who never hurts me because they never have the opportunity than I do with someone who has the opportunity but does not act on it, even if the empirical situation turns out to be such that the two people behave towards me in exactly the same way.[61]

In this way, mutual knowledge of what would happen in various counterfactual circumstances influences the character of relationships. The difference between stable co-operation based on a balance of power between rivals who are not strong enough to effectively dominate each other and a relationship of friendship consists in precisely this kind of counterfactual difference. In political society, common knowledge of the fact that other groups would modify the status quo in their favour were they to become strong enough to do so would have a deleterious effect on civic relations. By contrast, mutual knowledge that citizens are committed to justifying laws and policies on the basis of considerations that all reasonable citizens recognize as relevant would instantiate exactly the right kind of relationship between them, as it would bring about assurance that their shared commitments are not contingent on the balance of views held in their society.[62]

[60] Ibid.
[61] Ibid., 114.
[62] Ibid., 115-116.

C.

Though Lister provides a richer account of how the reasonable agreement principle is meant to instantiate relations of civic friendship, I think that ultimately it also fails as an answer to the rationale question for similar reasons to Ebbels-Duggan's view. Though many of the connections between relations of civic friendship and the reasonable agreement principle that he elaborates are plausible, it is far from clear that there are no other ways of instantiating civic friendship that do not appeal to the notion of reasonable agreement. And if there are other ways of having a valuable form of political community, then the value of civic friendship does not lend justificatory support to the reasonable agreement principle or its related duties on citizens and government officials in particular.

An argument to this end has been pressed by Paul Billingham, who suggests that the value of political community as specified by Lister could be realised without the reasonable agreement principle.[63] Suppose that instead of justifying political decisions by appeal to considerations that all reasonable citizens can be expected to accept, we have a democracy with the following features. Citizens deliberate about laws and policies openly and honestly, but by appeal to all of the reasons that they take to be relevant. There is no requirement to appeal to reasons that others can be expected to accept. Decisions are ultimately made using a democratic procedure of some kind, and citizens are permitted to vote on the basis of their assessment of the best reasons and arguments, again, without any requirement to vote on the basis of reasons that others can be expected to accept.[64] And the operation of this democratic procedure is constrained by a constitution which guarantees protection of the basic rights and liberties of all citizens.[65]

Billingham refers to this conception as 'argumentative democracy', and asks whether it could also realise the valuable form of political community that Lister argues we need the reasonable agreement principle for. If every citizen conscientiously engages with the arguments

[63] "Does political community require public reason? On Lister's defence of political liberalism," *Politics, Philosophy & Economics*, 15(1), 2016, 22-31.
[64] Ibid., 23-24.
[65] Ibid., 24.

offered by others, and takes each of their co-citizen's views to be equally valuable from the point of view of the democratic procedure, could this not realise a similarly valuable form of political community?

The fact that the citizens of an argumentative democracy share a commitment to democratic procedures as a way of making collective decisions seems to suffice to ensure that they do not see those decisions as aiming at alien values, which was Lister's first political-community-based reason in favour of the reasonable agreement principle. After all, argumentative democracy is also based on a set of shared commitments: to democratic procedures, to living together under the laws that are chosen by those procedures, to subjecting each other's views to critical scrutiny, and so on. And it is not clear why the different set of commitments that are shared under the reasonable agreement principle are needed to ensure that citizens do not see collective decisions as aiming at alien values.[66] Given the commitment to the protection of basic right and liberties argumentative democracy could not result in the state enforcement of a particular religious view, which would be starkest case of a collective decision that some would see as aiming at alien values. Of course, the decisions made in accordance with argumentative democracy may often have less in common with the views of some reasonable citizens than those made in accordance with the reasonable agreement principle. However, as Billingham rightly points out, Lister and others do not take civic friendship to be concerned with how much citizens share, but rather with their commitment to acting together in spite of their disagreements.[67]

Similarly, the shared commitments of argumentative democracy are such that mutual knowledge of their acceptance would suffice to ensure that citizens relationships are not degraded by an understanding of what might happen in various counterfactual scenarios. Again the commitment to basic rights and liberties means that even if the balance of views in society were to shift this would not lead to the erosion of those rights and liberties. The specific laws and policies that are chosen by the democratic procedure could of course fluctuate as the overall balance of

[66] Ibid., 25-26.
[67] Ibid., 26; Andrew Lister, *Public Reason and Political Community*, 108.

views in society changes, but such fluctuations would also occur under the reasonable agreement principle as citizens come to different views about the interpretation and weighing of shared values.[68]

Lister's view therefore runs into roughly the same problem that Ebbels-Duggan's view encountered. In order for the civic friendship rationale to be capable of justifying the reasonable agreement principle, it must be the case that in order to realise this valuable kind of political community we need the reasonable agreement principle *in particular*. It is not sufficient for its proponents to show that this is one way among others to realise a valuable kind of relationship or avoid a dis-valuable kind. What the proponent of the civic friendship rationale needs then is an argument for the claim that there is a valuable kind of community that only the reasonable agreement principle can make possible.

D.

In a recent article, R.J. Leland and Han van Wietmarschen have offered an argument of precisely this kind. They begin by claiming that all plausible accounts of friendship require friends to have a non-instrumental concern for each other's interests.[69] Whatever else friendship involves, this is a core feature of it. But what distinguishes their conception of friendship from those that we have discussed so far is that they argue that friendship requires a shared conception of the good. They motivate this conclusion with the following case:

> Erica thinks that excellence in intellectual, artistic, and athletic pursuits is what is important in life, while Patrick believes that a good life requires strong family relationships. […] These conceptions of the good come into conflict: if Patrick is offered a scholarship to an excellent university far away from his family, then Erica may think he should go, while Patrick may judge that he should not.[70]

Suppose that, in spite of their disagreement, Erica and Patrick still have a non-prudential concern to benefit each other. In that case, Erica could either respond to this conflict by: (1) deferring to

[68] Billingham, "Does political community require public reason?" 27-28.
[69] "Political Liberalism and Political Community," *Journal of Moral Philosophy*, 14(2), 2017, 157-158.
[70] Ibid., 158.

Patrick's conception of the good, or (2) by insisting on her own conception. But neither of these responses would be fully adequate, Leland and Wietmarschen argue. The former, as a general policy, would "border on insincerity and become hard to square with being a good friend"; the latter (again, as a general policy) would amount to a kind of paternalism and thus jar with the idea of friendship.[71]

Friendship therefore requires Erica and Patrick to have at least a *partially* shared conception of the good, so they can avoid a general policy of responding to each other in these two less than adequate ways. This need not mean that they must agree on a conception of the good down to the specifics. Perhaps Erica could accept that what is good for her is not good for Patrick, and vice versa. Or similarly, perhaps each could accept that it is good for one to live in accordance with one's choices. And what is true for Erica and Patrick is true for friendship in general, and scales up to the case of civic friendship. For civic friendship centrally involves a non-prudential concern to "benefit one another through political means, on terms that each regards and genuinely advancing the interests on herself and her fellow citizens".[72]

With this picture of civic friendship before us, we can see that once again the fact that reasonable citizens disagree about many matters can be seen to be a significant threat to the possibility of our being friends in this way. For as with Erica and Patrick, a policy of always deferring to the other person's judgment, or of always insisting on your own conception of the good would be inadequate. So, as for ordinary friendship, for civic friendship we need to have at least a partially shared conception of the good. But this somewhat trickier for a political community to achieve than it might be for individual friends. For we cannot suppose that all citizens in political society will converge on holding that a good life is one lived in accordance with one's own choices or any similar view: some reasonable conceptions of the good deny these claims.[73]

[71] Ibid.
[72] Ibid., 160.
[73] Ibid., 161, n. 33.

Fortunately, the reasonable agreement principle can help here. If each citizen deliberates on the basis of considerations that are shared by all reasonable citizens, then these decisions will ultimately be based on considerations that all accept. This amounts to sharing a conception of the good of a particular kind: a conception of each other's good *qua citizens*. By having a non-prudential concern for each other's interests couched in terms of this shared conception the good, citizens can avoid both the policy of consistently deferring to each other's judgment and the policy of consistently standing their ground in cases of where they have conflicting conceptions of the good. In this way, they can realise civic friendship in spite of their disagreements about the good life.[74]

How does all this avoid the problem faced by Lister's view? That problem, recall, was that it seemed perfectly possible to have the valuable form of political community he defended without the reasonable agreement principle. Citizens with a commitment to settling their disagreements via democratic procedures and living under the resultant laws could also constitute a political community. But if all citizens shared was a commitment to some basic rights and to settling their remaining disagreements via democratic procedures, then they would not *really* have a shared conception of good, Leland and Wietmarschen suggest: they would not be able to see the decisions that came out of the democratic process as aiming to further a conception of their shared interests.[75]

In order to evaluate this response to the problem, we need to ensure that we compare precisely what citizens share under the reasonable agreement principle as opposed to under argumentative democracy. In an argumentative democracy citizens do have *some* shared interests: in making their collective decisions via democratic procedures and in living under the resulting laws. However, Leland and Wietmarschen argue that these are shared interests in a 'very thin sense' and therefore considerations of civic friendship give us reasons to favour the reasonable agreement principle over argumentative democracy.[76]

---

[74] Ibid., 161.
[75] Ibid., 163.
[76] Ibid.

I think that in order to see what is wrong with this response, we need to return to the general framework that I set out at the outset of this chapter. We are considering the value or ideal of civic friendship as a potential answer to what I have called the rationale question: why should we accept the reasonable agreement principle? More specifically, given that there a variety of ways in which the reasonable citizens to which this principle refers could be idealised, why should we accept a particular version of the reasonable agreement principle, with its chosen level of idealisation? As Leland and Wietmarschen rightly point out, citizens have a thicker set of shared interests under the reasonable agreement principle than they do under argumentative democracy. Does that mean we have reason to favour the former over the latter? It would, if our aim were to ensure that citizens have the thickest possible set of shared interests. The thickest set of shared interests, of course, would be a single conception of the good that all citizens fully shared down to the details. But the starting point of the civic friendship rationale was that a community based on such a thick set of shared interests is impossible. Now all this may be quite right. Perhaps fully sharing a conception of the good down to the letter is too thick an ideal to be possible or desirable, and perhaps only sharing a commitment to living under democratic procedures is too thin an ideal to be valuable. However, in order to establish all this the proponent of the civic friendship rationale would need an explanation of what level of thickness is the right one. And Leland and Wietmarschen have not offered any such an explanation, beyond the claim that we need some shared interests to avoid the problems that faced Erica and Patrick. Perhaps this argument is better understood as holding that the ideal of civic friendship asks us to aim for the thickest level of shared interests that is possible subject to the constraint imposed by reasonable disagreement. If that is our aim, then there is a clear sense in which the reasonable agreement principle is preferable to argumentative democracy. But if the argument is understood in this way, what we will want to know is what kind of constraint reasonable disagreement imposes and why, for the rationale

question asks why we should accept this constraint. And on this rendering of the argument, the appeal to the ideal of civic friendship will not be able to answer this question.[77]

I have considered Larmore's view and the civic friendship rationale together in this chapter because they are each fairly self-contained attempts to defend the reasonable agreement principle. The same is not true for the two accounts I will discuss in the remainder of part I: those of Gaus and Rawls. I now turn to those views.

[77] I think that on this way of understanding the argument from civic friendship it would likely collapse into something like the argument from 'the limits of practical possibility' which I consider in chapter 4, section 4.5.

3. REASONABLE AGREEMENT AND THE REACTIVE ATTITUDES

3.1. Introduction

Gerald Gaus has been developing and defending a distinctive version of the reasonable agreement principle across a number of books and articles since the 1980s, and his view has now become one of the most discussed other than Rawls's.[78] Gaus offers sophisticated answers to the idealisation question, the rationale question, and the content question, and is acutely aware of the general difficulties that defences of the reasonable agreement principle face, such as those I raised in the previous chapter. Since his view has been modified in various was since it was first presented, I will draw primarily on the most recent full statement of it in *The Order of Public Reason*. Though in a few places I will also draw on his earlier works to ensure that we consider the best understanding of the view.

> Gaus's version of the reasonable agreement principle is the
>
> *Basic Principle of Public Justification*: A moral imperative "φ!" in context *C*, based on rule *L*, is an authoritative requirement of social morality only if each normal moral agent has sufficient reasons to (*a*) internalize rule *L*, (*b*) hold that *L* requires φ–type acts in circumstances *C*.[79]

In order to know when this principle is satisfied, we need to know when we have sufficient reasons to internalise rules and hold that they require particular acts.[80] Here Gaus defends

> *Having a Reason*: An agent has a sufficient reason *R* to hold that β is the thing to believe, or φ is the thing to do, if, and only if, (i) they have arrived at *R* by following the norms of good reasoning and (ii) if they engaged in a "respectable amount" of reasoning, they would not (or did not) discover defeaters for *R*.[81]

---

[78] This project begins with *Value and Justification: The Foundations of Liberal Theory*, (New York, NY: Cambridge University Press, 1990), continues with *Justificatory Liberalism: An Essay on Epistemology and Political* Theory, (New York, NY: Oxford University Press, 1996), and is given its most recent statement in *The Order of Public Reason.*

[79] *The Order of Public Reason*, 263.

[80] Internalise is used here as a technical term, defined as "having a motivational tendency of a particular kind to act" on the pattern of behaviour prescribed by the norm, where the kind of tendency in question is related to "a purpose of coordination". See Allan Gibbard, *Wise Choices, Apt Feelings: A Theory of Normative Judgment*, (Cambridge, MA: Harvard University Press, 1993), 68-71.

[81] *The Order of Public Reason,* 249-250.

Taking these claims together, we have an account of the justification of social-moral rules that takes seriously the evaluative standards of each individual. Social-moral rules are those "that require or prohibit action, and so ground moral imperatives that we direct to each other to engage in, or refrain from, certain lines of conduct."[82] In order for such a rule to be one others can legitimately demand my compliance with, it must cohere with my perspective in the right way—it must be one I have sufficient reason to endorse.

This gives us Gaus's answer to the idealisation question. Reasonable citizens are those who have followed norms of good reasoning and engaged in a respectable amount of reasoning following such norms. These idealised citizens will not have exactly the same beliefs and commitments as actual citizens, but there is a limit to how far their evaluative standards can stray from those of their un-idealised counterparts. For this reason, Gaus and his followers take their view to involve what they often call 'moderate idealisation'.[83] In keeping with the labelling of the previous chapter, I will refer to these moderately idealised agents as G-reasonable citizens.

This answer the idealisation question is similar to the one given by Larmore in that it also takes whether or not a particular view is reasonable to depend entirely on the procedure by which it came about. However, Gaus has more to say than Larmore did about what this procedure and its outcome must look like. He tells us that for a rule to be taken seriously as a candidate for passing the principle of public justification, it must satisfy the following conditions:

> *Generality*: Rules must be formulated in general descriptions rather than proper names or definite descriptions.
>
> *Weak Publicity*: Rules must be simple enough to be teachable to children.
>
> *Conflict Resolution*: Rules must order the competing claims of citizens, such that they can be appealed to in order to resolve such conflicts.

[82] Ibid., 2.
[83] Kevin Vallier, *Liberal Politics and Public Faith: Beyond Separation*, (New York, NY: Routledge, 2014), 145-180, for example.

Rules as Requirements: Rules must be such that they typically override citizens aims, values, and goals.

Reversibility: An individual's advocacy for a rule must not depend on the knowledge that she will only occupy specific roles or positions.

Modest Common Good: Rules must not undermine the core interests of other citizens.[84]

Each of these conditions is ultimately purely procedural. Even the modest common good requirement, which might appear to guarantee that the rules that satisfy the principle of public justification have some minimum guaranteed content is ultimately a procedural requirement.[85] Gaus holds that G-reasonable citizens will have different understandings of what is in the interests of others, and so the substantive content of the rules that satisfy the principle of public justification will have to be secured by showing that these moderately idealised agents would agree about what satisfies this requirement.

Given that this answer to the idealisation question retains the procedural character of Larmore's view, we might naturally think that it runs into the same problem: that there are no social-moral rules that all G-reasonable citizens will accept, and so the principle implies anarchism. However, Gaus also has an argument that attempts to avoid this conclusion. He employs what he calls an 'argument from abstraction', which aims to show that all G-reasonable citizens would affirm certain fundamental rights due to their necessity for effective agency.[86] That argument runs roughly as follows. Even in a pluralistic society, we cannot help but see ourselves as agents. Normal adult human persons are autarchic agents: they seem themselves as determining their own actions on the basis of their intentions and as forming their own intentions. On this basis, Gaus argues that his moderately idealised citizens would endorse various rights: to freedom of thought and conscience, and against coercion and deception, abstractly specified. However, they would not

[84] *The Order of Public Reason*, 294-303.
[85] Ibid., 302-303.
[86] Ibid., 334-369.

endorse a strong duty to assist others, as their evaluative standards may commit them to conceptions of justice that reject such duties.[87]

I have my doubts about this argument. Suppose our ordinary, un-idealised, citizen were a simple act-utilitarian, for example. It not clear why a sufficient amount of good reasoning in Gaus's sense, combined with the necessity of seeing herself as an agent would necessarily cause her to give up her utilitarianism and acknowledge the negative rights Gaus proposes. However, I will not be focussing my critique on this component of his view. Even supposing that this argument or a similar one can successfully show that Gaus's answer to the idealisation question can still permit him to give an acceptable answer to the content question, I believe his view would still be unacceptable on account of its answer to the rationale question.

What is Gaus's answer to the rationale question? The details of this will occupy us for the remainder of the chapter, but in brief he argues as follows. As moral persons we are liable to feel the reactive attitudes of resentment and indignation toward those who we perceive to violate moral rules, and these attitudes have appropriateness conditions. Suppose that I believe you have intentionally charged into me causing me to spill my cup of coffee, and I feel resentment on account of your lack of regard for me. If I were to learn that you were in fact caught in a strong gust of wind, my resentment would no longer be appropriate. The object of my attitude, your lack of regard for me, is no longer present, and therefore it can no longer be rationally sustained. Gaus argues that the reactive attitudes are appropriate toward ostensible rule-violators only if they had sufficient reason to internalise the rule in question.[88] He then defends the claim, following P.F. Strawson, that our propensity to the reactive attitudes is so deep a part of us that it sits beyond the reach of rational justification. Taking these claims together, he reaches the conclusion that there is a principle of public justification embedded in our social practice of issuing imperatives and following social-moral rules.[89]

[87] Ibid., 363.
[88] Ibid., 218-224.
[89] Ibid., 192.

In spite of the foundational importance of this argument for Gaus's defence of his principle of public justification, it has yet to receive sustained critical attention in the literature. In this chapter I evaluate and ultimately reject this, as I shall call it, reactive attitudes argument for the principle of public justification. I contend that—even granting some generous assumptions to his strategy—this argument fails.

3.2. The Reactive Attitudes Argument

The phrase 'reactive attitudes' is a term of art introduced by P.F. Strawson. These attitudes, Strawson tells us, are those that are responsive to our perception of the qualities of will that others express toward us as manifested in their actions or omissions.[90] In addition, these attitudes have 'vicarious analogues,' namely, the attitudes we adopt in response to our perception of the quality of another's will toward *others*.[91] Resentment is a paradigmatic instance of a reactive attitude. When I believe you have intentionally charged into to me causing me to spill my cup of coffee, my resentment is a response to my perception that your action expresses ill will toward me. When I learn that you were actually caught in a strong gust of wind, given that gusts of wind are not willed by anyone, I learn that my resentment is no longer appropriate, as my belief that you expressed ill will toward me is false. Perhaps some other attitude such as frustration would be appropriate, but not the distinctly *reactive* attitude of resentment.

There is some disagreement both about the range of attitudes that are reactive in Strawson's sense, and about what hangs them together as a class.[92] However, settling this issue is not essential for the reactive attitudes argument, which can be reconstructed as follows.

> (1) The reactive attitudes of resentment and indignation are appropriate only toward rule-violators who are i) capable of caring for moral rules, and ii) have sufficient reason to internalise the applicable moral rule.

---

[90] "Freedom and Resentment," in *Freedom and Resentment and Other Essays*, (Oxford: Routledge, 2008[1962]), 5-7.
[91] Ibid., 15.
[92] See, for example, Martin O'Neill, *Freedom, Fairness and Responsibility*, (Harvard PhD Dissertation, 2008), 77-83 and R.J. Wallace, *Responsibility and the Moral Sentiments*, (Cambridge, MA: Harvard University Press, 1996), chapter 2.

(2) Authoritative moral imperatives are appropriately addressed only to those to whom the reactive attitudes would be appropriate.

(3) Therefore, authoritative moral imperatives are appropriately addressed only to those who i) are capable of caring about moral rules, and ii) have sufficient reason to internalise the applicable moral rule.[93]

The first premise of the argument provides an account of when resentment and indignation are appropriate. As we saw in the spilt coffee example, certain beliefs are necessary to rationally sustain these attitudes. Discerning when resentment and indignation are appropriate is a matter of working out what the propositional content of these attitudes is. That is, answering the question: what must we believe about their object in order to sustain the attitude? Premise (1) expresses Gaus's answer to this question.

I will discuss (1) in more detail in section 3.4. Setting it aside for now, the controversial premise here is (2). Why should we believe that the fact that our reactive attitudes presuppose certain claims is sufficient to establish that those claims are correct? After all, an objector will be inclined to think, we can reflect on our attitudes and modify them—or perhaps even reject them completely—if we come to conclude they are unjustified. However, Gaus argues, following a strategy originally developed by Strawson, that the rational justification of these attitudes would be beside the point. The following passage most clearly expresses that strategy.

> We are embedded in certain sorts of practices, with certain beliefs and emotions. They form part of the reasons from which we must judge, criticize, and propose changes. A practice such as social morality is deeply embedded in our view of the world; it affects our understanding of interpersonal relations, including love and friendship, and so of what sort of life is worth living. If the presuppositions of our moral practices are so deep a part of the way we see the world, then to renounce the practice would be to renounce most of what we care for and value. But how can we have reason to do *that*? How could we survey all that matters to us and come to the conclusion that our reasons lead us to give it up, by renouncing the view of the world on which our reasons depends? Where would *that* reason come from? It is, I think, as difficult to argue a moral person out of her moral practices as it would be to argue the psychopath into them; given who they are, they do not have reasons to change their view of the world.[94]

[93] *The Order of Public Reason*, 205-258.
[94] Ibid., 192.

The inevitability of resentment and indignation for those bound up in the practice of experiencing them leaves them, in a sense, outside the reach of rational justification. If that is true, and we agree with Gaus about what the appropriateness conditions for these attitudes are, then we reach the conclusion. What is important here is the kind of conclusion we reach, and thus the kind of justification Gaus can say the principle of public justification enjoys. It is not "an exogenous (external) demand on an acceptable social morality based on some foundational moral intuition but a deep presupposition of our social morality with rational reactive attitudes".[95]

On introducing the conclusion that moral prescriptions are appropriately addressed only toward those who are capable of caring about moral rules and have sufficient reasons to endorse them, Gaus refers to it as *The Principle of Moral Autonomy*.[96] Its relevance for his principle of public justification should be relatively clear. If imperatives that demand we follow particular social moral rules should only be addressed toward those who have sufficient reason to internalise the rule in question, then we will want to know what rules (within the relevant jurisdiction) everyone has sufficient reason to internalise. This is what the principle of public justification tells us.

What does it mean to say that public justification is a deep presupposition of our social morality with rational reactive attitudes? Clearly Gaus does not think the principle of public justification is a conclusion we reach after extensive moral reflection, considering its implications and its coherence with our considered moral judgments about particular cases. Rather, it seems, he thinks it has normative force in the following way. We are, in virtue of the kind of beings we are—*moral persons*—committed to a certain kind of view of the world, including a certain framework of attitudes. To give up this framework is, to import some language from Strawson, practically inconceivable for us. Since, as premise (1) tells us, a presupposition of this framework of attitudes is *The Principle of Moral Autonomy*, we are committed to the idea of a publicly justified social morality in virtue of our commitment to this framework. It is the fact that our commitment to this framework is not something we can rationally reject that gives the conclusion its normative force.

[95] Ibid., 223.
[96] Ibid., 211.

There is evidence that confirms this understanding in the following passage from the preface to Gaus's earlier *Justificatory Liberalism*:

> We understand ourselves and others as capable of putting aside personal valuings, and of acting on norms that can be justified to all. To rid ourselves of this conception […] would undermine our understandings of ourselves, and others and social life, leaving us without rational grounding for most of what we hold dear. *Given who we are, […] we are committed to the idea of a publicly justified morality.*[97]

This completes my account of the reactive attitudes argument. It is undoubtedly an interesting argument for a principle of public justification, and considering whether it is successful will require taking a closer look at what Gaus and Strawson have to say about this strategy for defending premise (2), which I do in section 3.5. First, however, I will briefly consider the idea of *Having a reason* and what Gaus says in support of premise (1).

3.3. Having Reasons

Part of what makes Gaus's conception of public justification distinctive is his account of what it means to have a reason. Gaus presents us with an internalist, as opposed to an externalist, account of that relation. An externalist account would say that you have a reason *R* if and only if *R* is an external reason that applies to you.[98] Gaus rejects such views in the following passage.

> I believe the Externalist View of Having a Reason to be implausible: it misconstrues the relation between having a reason and being a rational agent. First consider the Externalist View of Having a Reason applied to theoretical reasoning. It implies that Aristotle, when writing on physics, had – possessed – a reason to embrace particle physics, because particle physics is true. But surely he did not have any such reason; to see R as a reason is to see it as justificatory, but Aristotle simply could not employ his rationality in a way that could lead him to see the facts supporting particle physics as justificatory (he could not even understand these facts). Only by not following the conclusions of his rational deliberation – being irrational – could Aristotle endorse such a "reason," and he could never see it as justifying a belief.[99]

[97] *Justificatory Liberalism*, vii, emphasis mine. This quote summarises an earlier formulation of the reactive attitudes argument from his *Value and Justification*.
[98] *The Order of Public Reason*, 233.
[99] Ibid., 233-234.

What does it mean to have a reason, then? We might suppose we stand in this having-relation to a reason *R* when it is a reason that applies to us, and we can be said to have or possess it.[100] However, consider a case of false belief, such as Bernard Williams's

> *Gin and tonic*: Alice orders a gin and tonic, but unbeknownst to her the bartender hands her a glass of petrol.[101]

Does Alice have a reason to take a sip in this case? If the glass had contained gin and tonic, we would naturally say that there is a reason for Alice to take a sip. She wants to drink a gin and tonic and it is a gin and tonic, so there is a reason for her to drink it. Further, she might sensibly be said to have this reason, as she quite justifiably believes it is a gin and tonic, given that that is what she ordered. In the case where the glass contains petrol, however, Alice may still have a warranted belief that it is a gin and tonic, but because this belief is false we would no longer say that there *is* a reason to drink it. Therefore, if we think we stand in this having-relation to a reason when it is both a reason that applies to us and one that we can be said to have or possess, we will say Alice does not have a reason to take a sip, as the reason in question does not exist.

On Gaus's view, however, whether Alice has a reason to take a sip does not depend on whether there is such a reason. In deciding whether we have a reason *R*, only our justified beliefs about whether *R* is a reason that applies to us are relevant, not whether there is in fact a reason *R*.[102] In considering *Having a reason*, then, it is important to be clear that what we are talking about is simply our warranted beliefs about the reasons that apply to us. With this in mind, nothing related to public justification follows directly from this having-relation. That an agent has a warranted belief that they should φ does not tell us whether we may permissibly stop them, or issue a morally authoritative imperative that they not take a sip. We need an explanation of why our warranted

---

[100] See, for example, Mark Schroeder, "Having Reasons," *Philosophical Studies*, 139(1), 2008. This is the factoring account of having a reason that Schroeder rejects.
[101] "Internal and External Reasons," in *Moral Luck: Philosophical Papers 1973-1980*, (Cambridge: Cambridge University Press, 1981), 102.
[102] *The Order of Public Reason*, 247.

beliefs about our reasons have normative import for such questions. This is what the conclusion of the reactive attitudes argument aims to provide, so let us now take a look at that argument.

3.4. The Appropriateness of Resentment and Indignation

As I noted in section 3.2, we discern when resentment and indignation are appropriate by working out what the propositional content of those attitudes is.[103] We also saw that Strawson claimed these reactive attitudes are a response to the quality of other's wills toward us, but this suggestion alone will not take us far. The pertinent question is: how should we specify the conditions that need to be met for an agent to count as expressing a relevantly malevolent will?

Gaus answers this question by defending two conditions, which I will now set out in turn. First, resentment is inappropriate toward those that are incapable of caring about moral rules when they do not promote their wants, ends, or goals.[104] They may care about rules when they correlate with their ends, such as when they know they can avoid an unpleasant punishment by conforming with the rule, but they lack the ability to understand that a moral demand can be authoritative, that it can require them to set aside what they care about and act in line with it. What they are lacking is in part an affective capacity. Not simply the ability to understand moral rules or reasons and their importance in cooperative social life, but the capacity to care about such rules.

This answer explains part of our everyday practice of holding people morally responsible. One thing we tend to take into account when deciding whether someone is responsible, and thus appropriately the object of resentment or indignation, is whether they possess certain capacities. It is not implausible, *prima facie*, to describe the relevant capacity as the capacity to care about and internalise moral rules. Young children, psychopaths, and gusts of wind do not have this capacity,

[103] Note that saying that these attitudes have propositional content is not to assume moral cognitivism. Rather, it is simply to say that these emotions are *about* something: that they have an object. The claim that resentment and indignation have a form of implicit demand as their object is common in discussions of the moral reactive attitudes. See, for example, Stephen Darwall, *The Second-person Standpoint*: *Morality, Respect and Accountability*, (Cambridge, MA: Harvard University Press, 2006). However, the claim that there is a strong link between the object of resentment and an action being wrong is controversial.

[104] *The Order of Public Reason*, 211.

but the kind of agents we generally take to be morally responsible do. I will refer to this first condition as the *capacity condition*.

This is not the only condition Gaus claims governs the appropriateness of resentment and indignation. He asks us imagine that we are issuing an imperative to someone unable to grasp its authority, in spite of the fact that they meet the capacity condition.

> I know that she does not see that it is the moral thing to do, and suppose I think her lack of appreciation is quite genuine. I demand "φ!" and she does not see why she is obligated to φ. She is puzzled that anyone would think φ is obligatory. If I think this, then again I cannot reasonably feel resentment or indignation that she fails to φ, any more than I can feel indignation at a four-year-old who is unable to detach himself from what he most wants to do and so steals some favourite candy. She just cannot see how "φ!" has any internal authority over her.[105]

He goes on to argue that if we offered this person a reason to adopt the rule that requires she φs, and she were unable to grasp the force of *that* reason, then resentment would still be inappropriate. She cannot grasp that φ-ing is the thing to do, because she cannot grasp my reason for holding that φ-ing is the thing to do. Therefore, I cannot rationally resent her failure to act.[106] I will refer to this second condition as the *sufficient reason condition*.

On this view, the capacity condition is what we can call a global condition. Where it is not met it tells us that due to some fact about the agent it is not appropriate to hold them responsible *in general*, and therefore that it would be inappropriate to resent them. The sufficient reason condition, by contrast, is what we can call a local condition. It tells us that due to some fact about the relationship between the agent and this particular event, it is not appropriate to hold them responsible for it, and therefore that it would be inappropriate to resent them. If you do not meet the sufficient reason condition with regard to some particular rule, you are still considered to be a responsible moral agent in general, who might be appropriately resented in other instances. If you fail to meet the capacity condition, however, you are not considered to be a responsible moral agent at all.

---

[105] Ibid., 219.
[106] Ibid., 220.

Taken together, these two conditions have a certain unity. The explanation of why some people can be the object of resentment and indignation is continuous with the explanation of why some specific actions or omissions can be appropriately resented. The explanatory factor in both the global case and the local case is the idea of having sufficient reason to internalise and comply with social-moral rules. If we accept this picture, the propositional content of these attitudes—*what* we resent, or feel indignant about, when we do so rationally—is the attitude of the violator's failure to follow a rule that they had sufficient reason to. When we feel indignant because we believe someone has violated a rule, global conditions such as insanity, or being a young child, and local conditions, such as being caught in a strong gust of wind, function to tell us that this attitude was not in fact present, and so our indignation cannot be rationally maintained.

However, it is far from obvious that resentment or indignation are never appropriate responses to those who do not meet the sufficient reason condition with respect to some rule. Intuitively, conscientious SS officers often acted wrongly and were thus appropriately resented, even if they would believe after a sufficient amount of good reasoning that any rules prohibiting their acts were ones they did not have sufficient reason to follow.[107] We might think, therefore, that we should accept an alternative view about when resentment and indignation are appropriate that does not have this implication. This is simply to doubt the normative significance of Gaus's notion of *Having a reason* for the question of when the reactive attitudes are appropriate.

One way to respond to this doubt would be to appeal to a deeper explanation of why it is inappropriate to resent those who do not meet the sufficient reason condition. However, as we shall see, Gaus's Strawsonian strategy precludes his providing such an explanation. That strategy is to claim that renunciation of the framework of reactive attitudes is practically inconceivable, and even if it were conceivable, it would be irrational. It tells that these attitudes lie beyond the scope of rational justification—so a deeper explanation of the conditions force is neither necessary nor possible.

[107] I adapt this example from Derek Parfit, *On What Matters: Volume One*, 158.

3.5. The Strawsonian Strategy

The preceding section set out what Gaus says in support of premise (1)—his account of when resentment and indignation are appropriate. In spite of the doubt I have just raised, for the sake of argument I will assume for the time being that, as a description of our current practice of holding one another responsible with its attendant liability to the reactive attitudes, this account is correct. That is, I will grant Gaus premise (1) of the reactive attitudes argument.

Attending to the second premise of that argument, recall that it states

(2) Authoritative moral imperatives are appropriately addressed only to those to whom the reactive attitudes would be appropriate.

The Strawsonian strategy that supports this premise is a much-discussed argument in the debate about the compatibility of moral responsibility and determinism. The stated aim of Strawson's "Freedom and Resentment," the paper that introduced this strategy, is reconciliation between the positions of the "optimist" and the "pessimist".[108] The pessimist holds that the truth of determinism would undermine our everyday practice of holding people morally responsible. Here is an example of how that practice might be undermined. One potential explanation of why we let people off the moral responsibility hook, and thus do not feel resentment on account of their actions or omissions that would otherwise be resentment-worthy is that they could not have acted otherwise. On this explanation, known as the principle of alternate possibilities, the fact that you were caught in a strong gust of wind gets you off the moral responsibility hook because it tells me you could not have acted other than to make me spill my coffee.[109] However, if this were the best explanation of why you ought to be let off the hook in this instance, the truth of determinism would render the practice of holding people morally responsible unjustified across the board. That determinism is true tells me that everyone is relevantly like you, unable to do otherwise.[110] The

[108] "Freedom and Resentment," 1-2.
[109] Harry Frankfurt, "Alternative Possibilities and Moral Responsibility," *The Journal of Philosophy* 66(23), 1969, 829.
[110] Note that I do not say what the thesis of determinism says here, as Strawson's argument proceeds without a precise definition. He instead proceeds on the basis that 'if there is a coherent thesis of determinism, then there must be a

optimist, by contrast, holds that our practice of holding people morally responsible is justified by its effectiveness at regulating social behaviour in ways that produce good consequences. According to this optimist then, the truth of determinism does not show our practice to be unjustified, since it does not tell us that our practice has negative consequences.

Strawson aims to reconcile these positions by showing they are both mistaken about the grounds of the practice of moral responsibility. On his view, while the grounds of the practice are not undermined by the truth of determinism as the pessimist thinks, the practice also is not justified by its good consequences, as the optimist thinks. His argument rests on an account of the reactive attitudes, and the conditions under which we consider it appropriate to suspend them.

One instance where we deem it appropriate to suspend the reactive attitudes, Strawson tells us, is where we learn that we were mistaken about the quality of will in question.[111] Again, the spilt coffee case is an example of this. When I learn that you were caught in a strong gust of wind, I learn that I was mistaken to hold that your spilling of my coffee expressed ill will toward me—it expressed no such will at all. This is what I called a local condition above. A second instance where we deem it appropriate to suspend the reactive attitudes is where we learn that it would be a mistake to resent this person due to a more general fact about them—their insanity, for example, or their being a young child.[112] This is what I called a global condition above.

Finally, beyond these two kinds of case, Strawson notes a further kind. We can, it seems, suspend the reactive attitudes not because, for one of the two preceding reasons, we believe it would be *inappropriate* to cling on to them, but rather because we have the ability to suspend them, which we can employ "as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity".[113]

sense of 'determined' such that, if that thesis is true, then all behaviour whatever is determined in that sense.' "Freedom and Resentment," 11.

[111] Ibid., 7-8.

[112] Ibid., 8-9.

[113] Ibid., 10.

Presented with this account of our practice of moral responsibility we will want to know what role it can play in answering the normative question of when we *should* suspend the reactive attitudes. After all, Strawson's pessimist presumably holds that regardless of when we *in fact* suspend the reactive attitudes, we ought to do so permanently, across the board, as the truth of determinism renders them unjustified.

To see Strawson's answer to this question we should first note that he holds that the alternative to the reactive attitudes is what he calls objectivity of attitude. He writes:

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided, though this gerundive is not peculiar to cases of objectivity of attitude.[114]

Importantly, this objectivity of attitude is fundamentally opposed to human interpersonal relations. To be involved in such relations precisely is, Strawson tells us, to be liable to the reactive attitudes.[115] The problem, then, is that that the pessimistic view just described would recommend we take up the objective stance permanently, as it tells us that our practice of holding people morally responsible is *always* unjustified. In taking up this stance permanently we would thus be sacrificing all interpersonal relations, and it is this that is

> for us as we are, *practically inconceivable*. The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question.[116]

Moreover, Strawson argues, even if we could conceive of having a choice in this matter, it would be irrational to choose to permanently adopt the objective stance. In making such a choice "we could choose rationally only in the light of an assessment of the gains or losses to human life; its enrichment or impoverishment; and the truth or falsity of determinism would not bear on the

[114] Ibid., 9.
[115] Ibid., 10.
[116] Ibid., 12.

rationality of *this* choice".[117] If he is right that the decision must be made on the basis of these considerations, then, taking into account the claim that the objective stance is fundamentally opposed to interpersonal relations, it is only a short step to the conclusion that permanent adoption of the objective stance is irrational.

So Strawson's answer to the question of what bearing an account of when we in fact suspend the reactive attitudes has on the normative question of when we ought to suspend those attitudes turns on what I will call the

> *Commitment claim:* Given our deep commitment to the reactive attitudes, it is practically inconceivable that we could abandon them; and further, if we could abandon them, it would be irrational to do so.

To be sure, this claim raises a number of further questions that would need to be answered in order to establish that it can provide a satisfactory answer to the normative question. As I noted above, this argument is now the subject of a large literature in debates about moral responsibility. I will not be able to offer anything close to an adequate analysis of it with respect to those debates here. What I will to do instead is grant, as a further assumption, that the commitment claim is true. Moreover, I want to grant that it provides a satisfactory answer to the pessimist's normative question. We *ought not* renounce the reactive attitudes in light of the truth of determinism, because our doing so is practically inconceivable in light of our commitment to interpersonal relationships. This is a controversial assumption. However, I will argue in the remainder of the chapter that even with it in place Gaus's argument for the principle of public justification is unsuccessful. Those who find this Strawsonian approach wanting more generally will simply have a further reason to reject that argument.

[117] Ibid., 14.

However, in granting the commitment claim as an assumption, I want to keep in the foreground why Strawson thinks this claim is true. It is because he holds that being involved in interpersonal relationships *precisely is* being liable to the reactive attitudes that he holds it to be practically inconceivable for us to give them up.

3.6. An Alternative

With this thought in mind, I will set out an alternative view of when the reactive attitudes are appropriate. Consideration of this alternative will, I believe, show that the scope of the commitment claim is limited.

As we have seen from the discussion so far, there are numerous cases where someone meets the capacity condition and ostensibly violates a moral rule, and yet we are inclined to think that it is not appropriate to feel resentment or indignation toward them, as in the spilt coffee example. As I noted above, if we think that the underlying principle that tells us when resentment or indignation are appropriate is the principle of alternate possibilities, then the compatibility of moral responsibility and determinism will be threatened. Since the truth of determinism will tell us that none of us could ever have acted otherwise, the principle will tell us that no one is ever the appropriate target for resentment or indignation.

An alternative principle, expounded by R.J. Wallace, is that of "no blameworthiness without fault".[118] On his view, our resentment or indignation can be rationally sustained where we believe a person has violated a moral obligation that we accept and hold them to. At first glance this may seem to have a limited ability to account for local conditions, because that I accept and hold you to a moral obligation seems like it would be unaffected by facts about your relationship to a particular act or omission. That is, it seems that if you are under a moral obligation to refrain from violating my bodily integrity by knocking into me, for example, the fact that you were caught in a strong gust of wind will be irrelevant to whether or not you have violated this obligation.

[118] *Responsibility and the Moral Sentiments*, 135.

However, Wallace argues that because holding someone to an obligation we accept involves a commitment to the existence of reasons that support that obligation, moral obligations must be focused on states of affairs that are directly susceptible to the influence of reasons. Therefore, an act that was not susceptible to the influence of reasons cannot constitute the violation of a moral obligation, and thus cannot make a person the object of rationally appropriate resentment.[119] This insight shows that the view can account for local conditions. When I learn you were caught in the gust of wind, what I learn is that your bodily movement was not the kind of thing that is susceptible to the influence of reasons, and therefore that you did not violate a moral obligation at all. Further, Wallace also supports a condition much like Gaus's capacity condition. It is not appropriate to hold someone responsible in general if they do not have the power of reflective self-control. That is, the power to grasp and apply moral reasons and regulate their behavior in light of them.[120] The underlying normative principle here, then, is that it is unfair to hold those who lack such powers responsible. Cases where we tend to let those with such powers off the hook are all cases where the agent turns out not to have performed that action intentionally at all. Importantly for Wallace the compatibility of moral responsibility with determinism is not threatened on this view, as even in a deterministic world we would have not have reason to suppose that no one ever has the power of reflective self-control, or so he argues.

I will call Wallace's alternative condition governing when we ought to let the generally responsible off the hook the *susceptibility to reason condition*. It should be clear that it is quite different from the sufficient reason condition. Provided Alf meets the capacity condition, he can act on his warranted beliefs about the reasons he has and still violate a moral obligation Betty holds him to, making her resentment appropriate. That is, even though he believes after a sufficient amount of good reasoning that he has a moral permission to φ, if he φ's, where φ is both an act susceptible to the influence of reasons, and an act Betty holds him to an obligation not to perform, her

[119] Ibid., 118-147.
[120] Ibid., 154-155.

resentment is rationally appropriate. On this view the propositional content of resentment is the belief that someone has violated an obligation we hold them to, where that requires, necessarily, that their act or omission was susceptible to the influence of reasons.

I have assumed that Gaus's account of when the reactive attitudes are appropriate in our current practice of moral responsibility is correct. I have granted, then, that within this practice local conditions are governed by the sufficient reason condition, not the susceptibility to reason condition. But supposing that is right, and we are currently participants in a practice governed in part by the sufficient reason condition, might we abandon that condition in favour of the susceptibility to reason condition? That would involve ceasing to suspend our resentment or indignation in the cases in which Gaus thinks we should, those in which we learn that the person in question did not have sufficient reason to internalise the rule we believed they had violated. Instead, we would suspend our resentment or indignation only in cases where we learn that either the person in question does not meet the capacity condition or their act or omission did not meet the susceptibility to reason condition.

In answering this question about the possibility of modifying our practice it seems that Gaus would want to appeal to the commitment claim. The sufficient reason condition is a presupposition of the reactive attitudes as we find them when examine our current practice. Given that the reactive attitudes are constitutive of interpersonal relationships, and we are committed to such relationships, we are committed to the reactive attitudes *as we find them*. However, this answer to our question is unsatisfactory. While it follows from our commitment to interpersonal relationships that wholesale repudiation of the reactive attitudes is practically inconceivable; it does not clearly follow that revising the conditions under which we deem it appropriate to suspend them is practically inconceivable. To defend that claim we would need to argue that this revision would amount to sacrificing all interpersonal relations, as Strawson thinks would be the case with permanent adoption of the objective stance.

Is there any reason to suppose that revising the conditions under which we deem it appropriate to suspend the reactive attitudes would have such severe consequences? Gaus appears to think this is the case. As I pointed out above, he tells us that to reject the understanding of ourselves and others as capable of acting on rules that can be publicly justified would be to "undermine our understandings of ourselves, and others and social life, leaving us without rational grounding for most of what we hold dear".[121] Further, he tells us: "If the presuppositions of our moral practices are so deep a part of the way we see the world, then to renounce the practice would be to renounce most of what we care for and value".[122] These claims echo Strawson's point about changing our social world so that there are no longer such things as the reactive attitudes, and therefore no longer such things as interpersonal relationships. However, it is difficult to see how, in Gaus's case, these claims could be true.

When the conditions under which we deem it appropriate to suspend the reactive attitudes are governed by the capacity condition and the sufficient reason condition, we have an expectation of others that they follow the rules that they have sufficient reason to internalise and comply with. When we feel indignant on account of their failure to do so, the object of our indignation is a fact about the relationship between their capacities and their act or omission. The object of our indignation is: "This person has the ability to care about and internalise social-moral rules, and they have sufficient reason to internalise and comply with this rule in particular—but they have violated it nonetheless!" By contrast, when the conditions under which we deem it appropriate to suspend the reactive attitudes are governed by the capacity condition and the susceptibility to reason condition, we have an expectation of others that they not violate moral obligations we accept and hold them to. The object our indignation, when we feel indignant on account of their failure to do so, is again a fact about the relationship between their capacities and the act or omission. It is: "This person's action, which violated a moral obligation, was susceptible to the influence of reasons

[121] *Justificatory Liberalism,* vii.
[122] *The Order of Public Reason*, 192.

and so can properly be seen to express a judgment on behalf of their object. They have *chosen* to φ, and given that φ violates an obligation I hold them to, I resent their choice to φ."

Both of these accounts involve demands and expectations. Both have an object that reflects a fact about the relationship between the person's capacities and the act or omission we might take to be worthy of say resentment, indignation, or gratitude. There might be numerous theoretical reasons to prefer one account to the other (or indeed to reject both) either as an account of what our practice is or ought to be. However, what seems clearly false is that Gaus can respond to the question of whether we ought to revise our practice from one to the other by appealing to the commitment claim to say that we should not, because that change would amount to renouncing all we care for and value. With this in mind we can see that, even granting the commitment claim, its scope is limited—it does not tell us that renouncing the sufficient reason condition in favour of an alternative is practically inconceivable for us.

Before moving on to reconsider the reactive attitudes argument in light of this, it is worth pausing to consider one way in which Gaus might reply to this argument. The susceptibility to reason condition seems to fit more intuitively with an externalist account of reasons and rationality. What makes that condition distinct from Gaus's sufficient reason condition is that it does not take the question of whether resentment is appropriate to hang on whether or not their target had sufficient reason, in his internalist sense, to endorse the rule in question. It takes the fact that there is a moral obligation not to φ, alongside the fact that the agent possesses the relevant capacity, as sufficient to ground rational resentment. This only makes sense, we might think, if we take external reasons to be normatively significant, such that an agent is guilty of a rational flaw for failing to respond appropriately to a reason that applies to them, even if they would not see it as a reason after a sufficient amount of good reasoning. Otherwise it would imply, somewhat counterintuitively, that we can rationally resent someone who is guilty of no rational failing at all. Since, as we have seen, Gaus rejects the externalist account of having a reason, we might think that he could thus reject any conditions that draw on externalism on the same grounds.

However, while Gaus rejects externalism as an account of the reasons that we *have*, he explicitly does not deny externalism about the reasons that there *are*.[123] We can happily accept that he is right to think that his internalism is the right account of the having-relation. The important question is whether or not that relation is of normative significance. As I noted at the end of section 3.3, Gaus thinks this relation is significant because of the reactive attitudes argument. It is because we only resent those who have sufficient internal reasons to endorse the rules they have violated—and we are committed to doing so in the sense implied by the commitment claim—that the reasons we have are significant for the question of when resentment is appropriate. Since in presenting the susceptibility to reason condition my aim has been precisely to call into question the nature of this commitment, Gaus cannot respond by simply asserting it.

3.7. Reconsidering our commitment to public justification

What are the implications of this argument about the scope of the commitment claim? It shows that the claim does not pick out the attitude of resentment with the sufficient reason condition as one of its appropriateness conditions in particular. Rather, it picks out a set of potential appropriateness conditions for resentment.[124] However, the reactive attitudes argument cannot support the principle of public justification unless the commitment claim picks out resentment with the sufficient reason condition as one of its appropriateness conditions in particular. To show this I will now reconstruct that argument taking the proper scope of the claim into account.

Premise (1) stated that resentment and indignation are only appropriate toward rule-violators who are (i) capable of caring for moral rules, and (ii) have sufficient reason to internalise the applicable moral rule. Our confidence in this premise will surely drop when we realize that the commitment claim does not provide support for it. We might therefore be inclined to think that (1) is not an accurate description of our practice of holding people responsible. If premise (1) is

[123] Ibid., 232-233.

[124] Some may prefer to refer to this alternative as an alternative attitude to resentment, rather than an alternative appropriateness condition for resentment. Nothing in the following argument hangs on this choice, but I refer to it as an alternative appropriateness condition for simplicity.

false, then Gaus's principle of public justification will not follow from the argument. However, I will now show that the principle does not follow even if we continue to grant this premise as an assumption.

We can begin by noting that premise (2) must be modified. Originally, it stated that authoritative moral imperatives are appropriately addressed only toward those to whom the reactive attitudes would be appropriate. But when we ask why this is true Gaus's response is to appeal to the commitment claim. As I have shown, this claim is only plausible with regard to a set of presuppositions of resentment, and at least one member of this set does not have the sufficient reason condition as a presupposition. The argument for this is simply the argument of the previous section: the susceptibility to reason condition is capable of playing the role of rationally grounding valuable interpersonal relationships. The availability of this alternative shows that the commitment claim does not establish that the sufficient reason condition is beyond the reach of rational justification. Rather, it shows that it would be practically inconceivable, and irrational if conceivable, to not endorse a member of the set of appropriateness conditions that is capable of furnishing rational grounds for human interpersonal relationships. Moreover, the commitment claim has nothing to say about how we choose from within this set—it only tells us that we cannot rationally choose to not to endorse any member of it. We must therefore modify (2) to state: authoritative moral imperatives are appropriately addressed only toward those to whom the reactive attitudes would be appropriate, according to an account of when these attitudes are appropriate. Where 'an account' can refer to any conception that is capable of furnishing rational grounds for these attitudes. The scope of appropriate moral authority, according to this premise, is dependent on the scope of the set of appropriateness conditions for the reactive attitudes that we cannot rationally reject.

Once the premise is modified in this way, the conclusion no longer follows. The conclusion, recall, was that authoritative moral imperatives are appropriately addressed only toward those who are (i) capable of caring for moral rules, and (ii) have sufficient reason to internalise the

applicable moral rule. This conclusion shows, Gaus claims, that public justification is a "deep presumption of our social morality with rational reactive attitudes".[125] However, the way in which (2) gives this conclusion its normative force in the original argument is, we saw, via the commitment claim. Since that claim does not tell us that we cannot modify the practice of social morality such that the sufficient reason condition is abandoned in favour of the susceptibility to reason condition, it does not give the conclusion its normative force in the modified argument. For all that has been said, we could adopt the susceptibility to reason condition, modify our practice as it recommends, and still enjoy rationally grounded interpersonal relationships without a commitment to a public justified social morality. Note that I need not even argue that we ought to do this. The mere availability of an alternative that is not ruled out by the claim strips the public justification principle of its normative force. On this reconstructed version of the argument, we are not committed to public justification in virtue of being the kind of beings that we are. Rather we are committed to rationally grounded interpersonal relationships. Public justification is one way, among others, that we can furnish rational grounds for these relationships. Therefore, public justification is not, as Gaus claims, a deep presupposition of our social morality with rationally grounded reactive attitudes, something that we cannot rationally reject given the kinds of beings that we are.

3.8 Conclusion

I will now summarise the argument I have developed over the course of this chapter. Gaus's argument for his principle of public justification appeals to the presuppositions of our framework of reactive attitudes. That argument aims to show us that we are committed to public justification by virtue of our commitment to this social-moral framework of attitudes. Some may find Gaus's description of our social-moral practice un-compelling, and still more may find his Strawsonian strategy wanting. However, I have argued that even assuming that Gaus is right on both of these points, the reactive attitudes argument fails. When we properly delineate the scope of the

[125] Ibid., 223.

commitment claim we see that our commitment is not to public justification but to interpersonal relationships, for which we can furnish alternative rational grounds. Perhaps other arguments can be marshalled in defence of the principle of public justification, but it is hard to see how they could establish that the principle enjoys the kind of justification Gaus claims for it. As I noted in the introduction to this chapter, Gaus's version of the reasonable agreement principle may well face serious difficulties when it comes to showing that there is anything that reasonable citizens, as he understands them, would unanimously accept. This is because, like Lamore, he takes reasonable citizens to be a constituency that is idealised only procedurally. However, if the argument of this chapter succeeds, then regardless of whether Gaus can avoid these difficulties, his view ought to be rejected. His answer the rationale question, that we ought to endorse the reasonable agreement principle because it is presupposed our moral reactive attitudes, is inadequate. Given this, I will now set his account aside and move on to discuss the version of the reasonable agreement principle defended by Rawls and his followers.

4. RAWLSIAN POLITICAL LIBERALISM

4.1. Introduction

The philosopher most commonly associated with the reasonable agreement principle is John Rawls. Though his views will occupy me to some degree throughout the thesis from this point on, this chapter will be devoted most specifically to them. I will begin by laying out in summary form what Rawls says in answer to the idealisation question, the rationale question, and the content question. Following this, I will consider some of the most well worked out attempts to argue that his answers to these questions can be worked up into a plausible and compelling defence of the principle. Approaching Rawls's view in this way is necessary, I think, because there is not a single uncontroversial interpretation of his position.[126] For this reason, the best way to approach Rawls's position is to examine both what he has to say in answer to the questions, as well as some of the best ways of interpreting or reformulating these ideas.

4.2. Rawls on the Idealisation Question

Over the course of *Political Liberalism,* Rawls can be understood as making the following set of claims about the attributes of reasonable citizens.

1. They (a) possess the two moral powers - the capacities for a sense of justice and for a conception of the good; (b) possess the intellectual powers of judgment, thought, and inference; (c) have a determinate conception of the good interpreted in the light of some comprehensive view; (d) are able to be normal, fully cooperating members of society over a complete life (81, 48-54, and 15-35);
2. They have the capacity to develop a reasonable moral psychology (81-86);

[126] It has been claimed, for example, that changes made between the hardback and paperback editions of *Political Liberalism* suggest a different kind of view. See Gerald Gaus, "The Turn to a Political Liberalism," in *A Companion to Rawls*, (Oxford: Wiley-Blackwell, 2014), 235-249, especially his distinction between the 'shallow' and the 'deep' version of view.

3. They recognize the burdens of judgment (54-58);

4. They recognize the five essential elements of a conception of objectivity (110-12).[127]

These attributes require some explanation, though it should be clear already that they go beyond Larmore's conception, where reasonable citizens are simply those "thinking and conversing in good faith and applying, as best [they] can, the general capacities of reason which belong to every domain of enquiry", and Gaus's conception, where they are the moderately idealised counterparts of actual citizens.[128]

Let me start with the first claim. The capacity for a conception of the good is a rational capacity. It is the capacity to work out, as an exercise of judgment, a conception of what one takes to be the object of rational desire.[129] Such a conception, Rawls's tells us, is "an ordered family of final ends and aims which specifies a person's conception of what is of value in human life".[130]The capacity for a sense of justice, by contrast, is the capacity to understand and act from a conception of justice. Such a conception specifies the principles for assigning rights and duties, and for distributing the benefits and burdens of social cooperation.[131] When citizens accept and act from a conception of justice, they limit their pursuit of their conception of the good in accordance with its precepts. This description of the two capacities might be taken to imply that reasonable citizens could have a conception of the good that considers violating the rights of their co-citizens to be valuable, but that they then also recognise the independent force of principles of justice and therefore do not pursue this value. Though nothing hangs on this presently, I do not think that this is what Rawls had in mind. In having a sense of justice, citizens take a conception of justice to

[127] Page numbers from *Political Liberalism*. I borrow this list from Leif Wenar's "Political Liberalism: An Internal Critique," *Ethics*, 105(1), 1995, 37, with one modification: I take the claim that reasonable citizens are 'ready to propose and abide by fair terms of cooperation, given the assurance that others will also' to be a part of the claim that they have the capacity for a sense of justice. Wenar lists this as a separate claim, though he suggests the modification I make here himself (n. 7), and I believe Rawls's text supports it.
[128] Larmore, "Political Liberalism," 340; Gaus, *The Order of Public Reason*, 249-250.
[129] In Rawls's view, 'good' is used as a synonym of 'object of rational desire'. See his account of the good in *A Theory of* Justice, chapter VII.
[130] *Justice as Fairness: A Restatement*, (Cambridge, MA: The Belknap Press of Harvard University Press, 2001), 19.
[131] Rawls, *A Theory of Justice*, 5.

limit their conception of the good itself, not merely their pursuit of the good. So a citizen with a sense of justice would not have, alongside this, a conception of the good that valued acting unjustly.

The second claim—that reasonable citizens must recognise the 'burdens of judgment'—is important. Stated more precisely, the claim is that reasonable citizens must recognise a set of ways in which their judgment may be burdened, and accept the implications of these burdens for reasoning about how political power ought to be exercised. The full set of burdens that Rawls specifies are as follows.

a. The evidence—empirical and scientific—bearing on the case is conflicting and complex, and thus hard to assess and evaluate.
b. Even where we agree fully about the kinds of considerations that are relevant, we may disagree about their weight, and so arrive at different judgments.
c. To some extent all our concepts, and not only moral and political concepts, are vague and subject to hard cases; and this indeterminacy means that we must rely on judgment and interpretation (and on judgments about interpretations) within some range (not sharply specifiable) where reasonable persons may differ.
d. To some extent (how great we cannot tell) the way we assess evidence and weigh moral and political values is shaped by our total experience, our whole course of life up to now; and our total experiences must always differ. Thus, in a modern society with its numerous offices and positions, its various divisions of labor, its many social groups and their ethnic variety, citizens' total experiences are disparate enough for their judgments to diverge, at least to some degree, on many if not most cases of significant complexity.
e. Often there are different kinds of normative considerations of different force on both sides of an issue and it is difficult to make an overall assessment.
f. Finally, […] any system of social institutions is limited in the values it can admit so that some selection must be made from the full range of moral and political values that might be realized. This is because any system of social institutions has, as it were, a limited social

> space. In being forced to select among cherished values, or when we hold to several and must restrict each in view of the requirements of the others, we face great difficulties in setting priorities and making adjustments. Many hard decisions may appear to have no clear answer.[132]

The aim of this list of burdens is to explain the 'fact of reasonable pluralism'. This is the fact that citizens with the capacities specified by the first attribute of the reasonable discussed above do not fully converge in their judgments, even in favourable conditions, accepting a single doctrine as true.[133] This fact, if it is a fact, calls for an explanation, as it might naturally be expected that persons exercising their powers of judgment in favourable conditions would come to converge in their judgments, at least in the long run. The burdens of judgment offer an explanation of why they do not.

The third claim is that reasonable citizens have a capacity for a particular kind of moral psychology that can provide the basis of their motivation to comply with principles of justice. Rawls describes this capacity by distinguishing three types of desire. Object-dependent desires are desires for a particular state of the world to come about, such a desire for food or rest. More precisely, they are desires whose object can always be adequately described solely by reference to state of the world and nothing else.[134] By contrast, principle-dependent desires are desires which cannot be described as desires for a particular state of the world to come about, because they cannot be described without making some reference to principles.[135] The desire to take the most effective means to an end you have, for example, is, in part, a desire to follow a particular principle and therefore cannot be adequately described without making reference to that principle. Conception-dependent desires are like principle-dependent desires in that they cannot be adequately described as desires for a particular state of the world to come about. However, conception-dependent desires are those that can described by saying that the principles that we

---

[132] *Political Liberalism*, 56-7.
[133] Ibid., 55.
[134] Ibid., 82.
[135] Ibid., 83.

desire to act from can be seen as belonging to a "rational or reasonable conception, or a political ideal".[136] For example, when you have a desire to live up to an ideal of citizenship or personhood, and this entails desires to follow multiple principles that you see as belonging to that ideal, you can be said to have a conception-dependent desire which is a desire to be a particular kind of citizen. In having the capacity for a moral psychology that can provide the basis of their motivation to comply with principles of justice, reasonable citizens have the capacity to develop conception-dependent desire of exactly this kind.

The fourth claim is perhaps the least discussed of Rawls's claims about reasonable citizens. He tells us that reasonable citizens recognise the essential elements of a conception of objectivity, which are as follows.

I. A conception of objectivity must establish a framework of thought that makes it possible to make judgments and draw inferences on the basis of mutually recognized criteria and evidence.

II. The judgments that element (I) makes possible must aim at being correct ('true' or 'reasonable'). Thus, a conception of objectivity must specify the notion of a correct judgment made from its point of view.

III. A conception of objectivity must specify an 'order of reasons' given by its criteria. These reasons are to be assigned to agents, as reasons they ought to be guided by, such that they ought to act from these reasons whether or not they accept them.

IV. A conception of objectivity must take the point of view given by its criteria for correct judgment as distinct from the point of view of any particular agent or group of agents.

V. A conception of objectivity must have an account of agreement in judgment. Assuming agents reason correctly from the appropriate point view, and rely on the same information, their judgments will converge.[137]

[136] Ibid., 83-84.
[137] Ibid., 110-112.

What these elements amount to is that reasonable citizens must accept a conception of objectivity that distinguishes true or correct moral and political judgments from a variety of other phenomena, such as merely expressing a psychological state or trying rhetorically to persuade others. A conception of objectivity does this by defining an objective point of view, which is distinct from the point of view of any particular individual or group of individuals.[138]

These four claims complete what Rawls has to say about reasonable citizens specifically, but they do not complete his answer to the idealisation question. When I introduced the idealisation question I noted that in answering it we are not limited to making claims about the features of reasonable citizens—we can also make claims about the background conditions they are in. And Rawls's claim that completes his answer to the idealisation question is exactly this:

5. They are citizens of a well-ordered society.

A well-ordered society (WOS hereafter) is an idealised society that meets the following conditions. First, everyone accepts the same conception of justice. Second, everyone knows that everyone else accepts that same conception of justice. Third, this conception of justice effectively publicly orders the society. Fourth, the conception of justice and its grounds are public knowledge among citizens—they are aware of how it balances their competing claims and the arguments that can be given in its favour.[139]

4.3. Rawls on the Rationale Question

I will call citizens who are idealised in all of these ways R-reasonable citizens. Why should we care about what they would agree to? Rawls tells us that the principal question he aims to address in *Political Liberalism* is

> How is it possible for there to exist over time a just and stable society of free and equal citizens who remain profoundly divided by reasonable religious, philosophical, and moral doctrines?[140]

[138] The point Rawls wishes to express by saying that the objective point of view is distinct in this way is that a person's thinking something to be true alone never makes it true. What makes a particular true claim or correct is always that it is would be the outcome of correct reasoning according to the norms that define the objective point of view. Ibid., 111.
[139] *Justice as Fairness: A Restatement*, 8-10.
[140] *Political Liberalism*, xxxvii.

What does it mean to say that a just society is also *stable*? In ordinary language, to refer to a society as stable is generally to say that it has a relatively harmonious social order that could not be too easily upset. Potential setbacks to this kind of stability might be things like the overthrow of a government by force, or a significant disruption to workings of the economy. But importantly, the extent of a society's stability in this sense looks like an empirical question. In order to determine whether or not a society was stable, or how stable it was, it seems the way to proceed would be to try to establish what the empirical determinants of a harmonious social order are, and to work out to what extent these are realised in a given society. It is certainly not the kind of question that can be answered by philosophical analysis from the armchair.

In Rawls's mouth, by contrast, stability is quite distinct from the property just described. The type of stability to which the question refers is 'inherent stability', or 'stability for the right reasons'. [141]This is a property that might be possessed not by societies as we know them, but rather by societies that are well-ordered by a conception of justice.

In order to understand Rawls's answer to the rationale question, it is necessary to first understand what he means by stability for the right reasons. Rawls's defence of his conception of justice has always had two distinct stages. The aim of the first stage is to establish that parties to the original position (OP hereafter) would choose his conception of justice over various rival conceptions. The OP is a procedure in which deliberators are tasked with choosing principles of justice to publicly order a WOS. The deliberators in the OP are considered to be rational representatives of citizens of the WOS. However, they do not know the particular ends of the citizens that they represent. More specifically, the parties are situated behind the veil of ignorance, meaning they do not know the race, ethnicity, gender, age, income, wealth, natural endowments or conception of the good of the citizen that they represent. What the parties do know are general facts about human social life, including uncontroversial conclusions of science, economics and psychology, and that the person that they represent is instrumentally rational and possesses a

[141] *A Theory of Justice*, 436; *Political Liberalism*, 143.

determinate conception of the good.[142] Rawls argued that parties to the OP would choose two lexically ordered principles of justice:

> First, each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all.
> Second, social and economic inequalities are to satisfy two conditions: (a) they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and (b) they are to be to the greatest benefit of the least-advantaged members of society.[143]

Having argued that they would choose these two principles, Rawls aims in the second stage to establish that this conception of justice would be stable. The stability question, as he puts in *Theory* is: would this conception engender a sufficiently strong desire to do justice in these citizens that they would give it a regulative place in their rational plan?[144] That is, would they desire to be a just citizen—one who has a highest-order desire to act in accordance with the conception, and to rid herself of desires to act unjustly? This is, on its face, a psychological question about the ends of citizens would have if they had grown up under and been educated in a society well ordered by that conception of justice. If we can answer this question in the affirmative, then we can say it is rational for citizens in the conditions of a WOS to respond to the fact that others are acting justly by continuing to act justly themselves. One way to re-describe this conclusion is to say that stable WOS is in equilibrium: from the point of view of each citizen it is rational to continue to act justly.[145]

The question from the point of view of each of these citizens is: should I maintain my desire to be a just citizen? In asking this question, we are assuming that each of the members of the WOS already has this desire, and they are asking themselves what future it has in their rational plan.[146] This assumption does not rule out the possibility of their choosing to abandon it. A further assumption is that we are to conceive of the citizens as reasoning about this choice while setting

---

[142] *A Theory of Justice*, 102-130.
[143] This is Rawls's most up to date statement of the two principles, taken from *Justice as Fairness: A Restatement* (42) rather than from *Theory*.
[144] *A Theory of Justice*, 397.
[145] Ibid., 436; Paul Weithman, *Why Political Liberalism? On John Rawls's Political Turn*, 43-51.
[146] This element of the account is captured well by Anthony Laden's metaphor: "We can picture this by thinking of […] setting up toy models of different societies and setting them running." "Games, Fairness, and Rawls's *A Theory of Justice*," *Philosophy & Public Affairs*, 20(3), 1991, 212.

aside their desire to be just for its own sake. They are to ask (and this is the final and most precise presentation of their question): do I consider being a just citizen to be a regulative object of rational desire, when I set aside my desire to be just for its own sake and consider only my other desires? That is, would I choose to give this desire a regulative place in my rational plan if I bracketed the desire itself and reasoned on the basis of my other desires?[147]

This idea about bracketing or setting aside the desire to be a just person needs some explanation. Consider the following example. Alf has made a promise to his friend Betty. He knows that justice requires him to keep that promise; however, he might hold that being a good friend also requires him to keep it. In such a case he could reason his way to the conclusion that he ought to act as a just citizen would, while bracketing his desire to be just and considering only his desire to be a good friend. This is possible here because the action of keeping the promise simultaneously satisfies two descriptions: 'acting justly' and 'being a good friend'. In order to hold that citizens of a WOS would desire to be a just person when reasoning in this way it must be argued that they would develop desires that appropriately corresponded with the content of justice, like Alf's desire to be a good friend. It must also be shown that they would not develop desires that pulled in the opposite direction; such as if Alf were to have a competing desire to renege on his promises when doing so was to his financial benefit.

Different kinds of forces could render it rational for citizens of the WOS to have a regulative desire to be a just citizen when they reason in this way. Suppose Betty has a desire not to be imprisoned. If the conception of justice that well ordered her society held that the punishment for breaking promises was imprisonment, she could reason her way to the conclusion that she ought to act as a just citizen would by considering only her desire not to be imprisoned. In fact, we might infer from the assumption that citizens of a WOS will want to avoid severe punishments, that if those who violated the principles of justice were likely to face sufficiently such punishments, a conception could be stabilised regardless of the content of its principle and ideals. Considering

[147] *A Theory of Justice*, 499.

her desire to not be imprisoned, Betty could also reason her way to the conclusion that she ought to follow principles with unappealing or seemingly arbitrary content, such as: 'harm my friends whenever I can get away with it,' or 'only keep promises on Tuesdays'. All that would be required is a sufficiently credible threat of punishment in conjunction with a sufficiently strong desire to avoid it.

The final component of this account of stability is a distinction between these different kinds of forces that could stabilise the WOS. A conception of justice is inherently stable if its stability can be brought about by the educative effects of living under institutions that satisfy the conception of justice in question. That is, the effects on our desires of growing up under particular institutions: those institutions that bring about a conception of distributive justice, publicise it, and are responsible for the moral education of citizens. If any other effects such as the use of a penal system or indoctrination into an ideology are required to stabilise a conception of justice in the WOS then that conception is not inherently stable—it may be stable, but its stability is imposed rather than inherent.[148] To return to our example, suppose Betty has a set of desires such that when she brackets her desire to be just and draws up a rational plan she chooses to maintain her regulative desire to be a just person. If all citizens of the WOS are like Betty in this regard, the conception of justice in question is stable. However, in order for it to be inherently stable it must also be the case that Betty would develop these desires in circumstances where no punishment was attached to failure to act in accordance with the conception of justice. In this case her choice to maintain her desire to be a just citizen stems from the free exercise of her capacity for reason.[149]

Rawls made a set of arguments in *Theory* that aimed to establish that a society well-ordered by justice as fairness would be stable in this way. He then famously came to believe that these arguments were unsuccessful, prompting the changes he made to his view in *Political Liberalism*.[150]

---

[148] This claim is subject to the caveat that the use of a penal system will be necessary to generate assurance among citizens of the WOS, Ibid., 211. I discuss this caveat in more detail in chapter 5.

[149] For a clear account of the distinction between these two kinds of stability, see Weithman, *Why Political Liberalism?* 43–51.

[150] For the claim that it was the failure of the stability arguments that prompted the changes, see *Political Liberalism*, xv.

However, for the time being it will not be necessary to go into the details of these arguments. In order to evaluate whether or this appeal to stability for the right reasons can successfully answer the rationale question, what we need to know first and foremost is why we ought to care about whether our conception of justice is stable in this way. I will consider this question in more detail shortly, after setting out Rawls's answer to the content question.

4.4. Rawls on the Content Question

So the rationale for justifying the exercise of political power to R-reasonable citizens is to show that a just WOS could be stable for the right reasons. What does Rawls think R-reasonable citizens would accept? His answer to the content question is that every reasonable citizen will accept a conception of justice that is 'political' and 'liberal'. A political conception of justice is a moral conception designed to apply solely to a particular subject matter, the basic structure of society. This is a society's main political, social, and economic institutions, and how they fit together to form a unified system of social co-operation. In saying that a political conception is designed to apply solely to this limited subject matter is to put it in contrast with conceptions that are general and comprehensive; those that have the entirety of human value as their scope.[151] A conception of justice is 'liberal' if it includes at least the following three features. First, ascribing to all citizens' familiar individual rights and liberties, such as rights to freedom of expression, liberty of conscience, and free choice of occupation. Second, giving special priority to these rights and liberties, especially over demands to further the general good or perfectionist values. Third, assuring for all citizens' sufficient all-purpose means to make effective use of their freedoms.[152]

For all I have said so far, it might seem surprising that Rawls's answer to the content question is not simply that his two principles of justice are what all R-reasonable citizens would accept. Part of the explanation of this fact is that his concern in defending the reasonable agreement principle is a concern not with solely justice but also with legitimacy, which he conceives of a

[151] Ibid., 11-14.
[152] Ibid., 6.

weaker standard than justice.[153] He defends the following principle of legitimacy, and a corresponding natural duty that falls on citizens.

> *Liberal Principle of Legitimacy*: Our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason.[154]
>
> *Duty of Civility*: We must be able to explain to each other, on fundamental questions, how the principles and policies we advocate can be supported by the political values of public reason.[155]

A conception of justice that is both political and liberal in the sense just described is one that can all R-reasonable citizens can be expected to accept, and so exercising political power in accordance with such a conception is legitimate according to the liberal principle of legitimacy. Justice as fairness is a conception that is political and liberal, but it is not the only such conception. So Rawls did not think that justice as fairness uniquely satisfied the reasonable agreement principle: he thought it was one of a family of conceptions of justice that could satisfy it.[156]

4.5. The Internal Conception of Political Liberalism

This summary of Rawls's view raises a number of questions that would need to be answered for it to provide a satisfactory defence of the reasonable agreement principle. As I mentioned above, there is no single uncontroversial interpretation of Rawls's view, so rather than trying to construct a plausible account of the principle from the above summary, I will examine some of the most significant interpretations that have been put forward by others. I begin with Jonathan Quong's view.

---

[153] Ibid., 428.
[154] Ibid., 137.
[155] Ibid., 217.
[156] Ibid., 167.

Quong draws a distinction between two ways in which we might understand Rawls's political liberalism: the external conception and the internal conception.[157] This distinction is fundamental to his answer to the idealisation question, so I will begin by describing it.

The external conception says the aim of political liberalism is to show how liberal principles can be justified to the actual citizens of liberal democratic societies. This aim leads it to a particular interpretation of the idea of reasonable disagreement, a particular answer to the idealisation question, and thus a particular account of the reasonable agreement principle. Let me start with the idea of reasonable disagreement. The external conception holds this to be a fact about the world which has implications for theories about how we ought to exercise political power. This is because, in answer to the idealisation question, it holds that reasonable citizens are a subset of the actual citizens we find in present day liberal democratic societies. However, this does not mean that the view holds that the exercise of political power must be justifiable to all actual citizens; rather, reasonable citizens are an idealised subset of the actual population.[158]

On the internal conception, the aim of political liberalism is not to show how liberal principles can be justified to the actual citizens of liberal democratic societies, but rather to establish the more modest conclusion that liberal principles can be justified to the citizens of a well-ordered society. The idea of reasonable disagreement, on the internal conception, is not a fact about the world but a fact about liberal theory itself. The view sees reasonable disagreement in this way because it takes reasonable citizens to be the citizens of a WOS, and any society well-ordered by liberal principles, Quong tells us, will be characterised by reasonable disagreement.[159]

Quong argues that the key flaw of the external conception is that it holds that liberal principles must be justified to 'persons who cannot be assumed to endorse any basic liberal norms or values,' something he thinks cannot be done.[160] This reveals an important detail about the internal conception's answer to the idealisation question: reasonable citizens are *assumed* to have a

[157] *Liberalism without Perfection*, 138-145.
[158] Ibid., 141.
[159] Ibid., 143.
[160] Ibid., 140 and 146.

commitment to some basic liberal ideas. More specifically, Quong holds that we are to understand reasonable citizens as being committed to the moral ideal that persons are free and equal, and to the ideal of society as a fair system of cooperation between such persons over time. In detailing what this commitment amounts to, he suggests that we understand reasonable citizens as being committed to the values of freedom, equality and fairness that are embodied in those ideals.[161] So, in short, Quong's answer to the idealisation question is: reasonable citizens are citizens with a commitment to the values of freedom, equality, and fairness. I will thus refer to persons idealised in this way as Q-reasonable citizens.

How about the rationale question? Why should we care about what Q-reasonable citizens do or do not accept? Quong states that the goal of the internal conception is

> to show that the kind of citizens that would be raised in a society well-ordered by a liberal conception of justice could endorse and support their own liberal institutions and principles if those institutions and principles were justified in a particular way. It shows that liberalism, when understood as political and not metaphysical, can generate its own support under ideal conditions, and thus is not incoherent or unstable.[162]

That is: in order to show that liberalism is not incoherent or unstable, we need to show that those raised in a society well-ordered by a liberal conception of justice could come to endorse it. But this answer just raises two difficulties.

First, it raises but does not answer the question of why we ought to care about the kind of stability or coherence that Quong is appealing to. If stability were the ordinary language property I discussed above, then it would perhaps be obvious why we ought to care about it. If enacting an unstable conception of justice would cause a breakdown of social order, then we could say that it should not be enacted because a harmonious social order is a precondition for a number of other valuable goods. But this is not what stability for the right reasons refers to, and as I said above it is an empirical question whether enacting a particular conception of justice would have this effect. A conception of justice is stable for the right reasons when R-reasonable citizens would freely choose to affirm it. This is also an empirical question to the extent that it requires an empirical judgment

[161] Ibid., 140.
[162] Ibid. 158.

about the desires that the citizens of the WOS would come to have, but it is not an empirical question about what would cause the breakdown of social order here and now. Similarly, it is not clear what the incoherence is meant to amount to. Other theories in moral and political philosophy do not think they need to show that the conception of justice they defend is stable for the right reasons. Perhaps they are making a mistake, but it is hard to see in what sense an appeal incoherence explains what they are doing wrong. A conception of justice that is not stable for the right reasons need not be incoherent in the sense of recommending we do both P and not P, so absent some further explanation it is hard to see what is wrong with views that do not establish this kind of stability. Further weight can be put behind this point by noting that in answering the rationale question the aim is to explain why acceptability to a particular idealised constituency is important. But the property of being stable for the right reasons *just is* the property of being acceptable to a particular idealised constituency. So there is no way an appeal to this kind of stability can answer the rationale question on its own. If it were obvious that we ought to care about this kind of stability, the rationale question would not even arise in the first place.

At this stage it is worth pointing out that there are at least three reasons why we might doubt that stability for the right reasons is a condition a conception of justice must satisfy. First, we might think that whether or not a conception of justice is capable of ordering a society in this way is simply irrelevant to the question of whether it is the true, or best justified, conception of justice.[163] G.A. Cohen argued it would be absurd to think principles of justice must be stable, as that would preclude us from saying such sensible things such as 'we don't want our society to be just only for the time being: we want its justice to last'.[164] To say that principles of justice are required to be stable is to confuse two separate questions: the question of what is just, and the question of what would be stable.

Second, we might think that the true, or best justified, conception of justice would *by its very nature* be unable to order a society in this way, as public knowledge of it would undermine its

[163] G.A. Cohen, *Rescuing Justice and Equality*, (Cambridge, MA: Harvard University Press, 2008), 327-31 and 364-68.
[164] Ibid., 328.

ability to achieve its aims. This view originates with Henry Sidgwick's utilitarianism in *The Methods of Ethics*.[165] Sidgwick did not think that we should encourage all ordinary citizens to follow the principle of utility in their everyday lives, as it was too likely that they would make mistaken calculations. Instead, we ought to keep the truth of the principle of utility from them, and encourage them to follow a simpler set of moral rules closer to those of common sense morality. This deception is warranted by utilitarianism itself, as such a strategy would be more likely to maximise utility. This is to explicitly deny that principles of justice need to be stable for the right reasons. After all, stability for the right reasons requires that all of the citizens of the WOS understand and comply with the principles of justice in their everyday lives.

Third, and finally, we might think that even if the ability to well order a society *is* a desideratum for a conception of justice, it is surely but one that must be weighed against others in the process of working out which is the true, or best justified conception of justice.[166] That is, we might think that a conception of justice's ability to generate its own support in ideal conditions is an important property, and it is an advantage if our conception has it. But it is also not the be all and end all. We should be willing to accept a little less stability (or likelihood of stability) in exchange for other theoretical virtues. On this view, though stability for the right reasons is a good property for a conception of justice to have, it is not a condition that must be satisfied for it to be plausible.

Setting aside these doubts about the significance of stability for now, the other issue is that Quong's answer to the idealisation question was not that reasonable citizens are those raised in a WOS, but rather those who accept the values of freedom, equality, and fairness. Showing that liberal principles can be justified to those with a commitment to these three values does not show that those raised in society well-ordered by a liberal conception of justice would come to endorse liberal principles—these are separate questions, unless we take as a given that those raised in a society well-ordered by a liberal conception of justice will accept the values of freedom, equality, and fairness. So even if we had a compelling explanation of the significance of stability, we would

[165] (Indianapolis, IN: Hackett, 1981[1907]), 489.
[166] Joseph Raz, "Facing Diversity: The Case of Epistemic Abstinence," *Philosophy & Public Affairs*, 19(1), 1990, 16.

not have reason to think that Quong's account had established that any liberal conception of justice would be stable.

This second issue is somewhat less severe than the first. Since Rawls provided an account of how citizens of the WOS would come to affirm justice as fairness which included a moral psychological account of the development of a sense of justice, Quong may simply be able to help himself to this account.[167] If Rawls's arguments did successfully establish that the members of the WOS would affirm justice as fairness, then it seems that *a fortiori* those arguments could be used to establish that they would affirm the values of freedom, equality, and fairness. Justice as fairness is, after all, a conception of justice that provides an expression of those values. However, if Quong wants to answer the rationale question by appealing to stability for the right reasons, he cannot avoid defending either this moral psychological account or a related one. One reason the stability answer to the rationale question might be doubted is scepticism that citizens raised in a WOS would actually come to affirm these values, and the only way to address such doubts would be to provide some reasons to think that they would.[168]

The first issue is serious, however. In order to successfully appeal to stability for the right reasons in answer to the rationale question, we need an explanation of the significance of stability. Why might this kind of stability be important? Quong appeals, somewhat briefly, to the idea of 'the limits of practical possibility' in answer to this question.[169] I will now explain this idea and consider whether it can offer a compelling explanation.

Rawls claimed that one role of political philosophy is to probe the limits of practical possibility, by setting out a realistically utopian ideal of social cooperation.[170] The rationale for wanting to probe these limits is relatively straightforward. If we go beyond the limits of what is practically possible and propose an ideal that cannot be achieved, then, given 'ought implies can',

[167] *Theory*, chapter VIII.
[168] This paragraph draws on Paul Billingham's "Liberal Perfectionism and Quong's Internal Conception of Political Liberalism," *Social Theory and Practice*, 43(1), 2017, 104.
[169] *Liberalism without Perfection*, 158-159.
[170] *Justice as Fairness: A Restatement*, 4-5.

we will not have succeeding in setting out an ideal that gives us genuine guidance about what we ought to do. However, if we recommend an ideal that falls short of the limits of practical possibility, then we may have left a more attractive possibility on the table, and thus our recommendation will fall short in this way.[171]

Since neither Rawls nor Quong have specified in detail the relationship between the limits of practical possibility and stability for the right reasons, some reconstruction will be necessary to work out how the former might explain the significance of the latter. But at first pass, the explanation looks promising. Given the reasons I have just suggested as to why one might doubt the significance of stability, a satisfactory explanation of its significance must explain why we ought to want our conception justice to be stable, and further why this is a requirement it must meet, rather than merely a desirable property. An appeal to the limits of practical possibility looks like it has the potential to do this explanatory work, provided we think it is a requirement that our conception of justice is practically possible.

Before analysing this answer, let me first take stock of where we are and what exactly it is that this appeal to the limits of practical possibility would need to explain. On the version of the reasonable agreement principle under consideration, the answer to the idealisation question is that reasonable citizens are the citizens of a well-ordered society. The question we are now trying to answer is the rationale question: *why* must the exercise of political power be justified to the citizens of a well-ordered society? On the answer under consideration, the exercise of political power must be justified to them because this establishes that our conception of justice is stable for the right reasons. However, I have argued that, for various reasons, this appeal to stability *alone* cannot provide a satisfactory answer to the rationale question. With this in mind, we are now considering whether the significance of stability can be established by appealing to the thought that our conception of justice ought to be realistically utopian, or at the limits of practical possibility.

[171] I follow Mark Jensen's "The Limits of Practical Possibility," *The Journal of Political Philosophy*, 17(2), 2009, 168.

I think a significant difficulty with this proposal can be drawn out by taking the view of the utilitarian objector to stability for the right reasons as a foil. Stability for the right reasons requires, we have seen, that a conception of justice is generally and freely accepted in the circumstances of a WOS. A proponent of the principle of utility, however, may deny that it is necessary or desirable that their principle be generally accepted by citizens. The most influential proponent of this view was Sidgwick, as I meantioned earlier, and it is now worth quoting him in full:

> The Utilitarian conclusion, carefully stated, would seem to be this; that the opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric. Or, if this concealment be difficult to maintain, it may be desirable that Common Sense should repudiate the doctrines which it is expedient to confine to an enlightened few. And thus a Utilitarian may reasonably desire, *on Utilitarian principles*, that some of his conclusions should be rejected by mankind generally; or even that the vulgar should keep aloof from his system as a whole, in so far as the inevitable indefiniteness and complexity of its calculations render it likely to lead to bad results in their hands.[172]

Sigwick's claim is that since it may be desirable according to the principle of utility for most citizens to reject utilitarianism, the principle itself may direct us to conceal its truth from them. This claim has been subject to considerable criticism, most notably by Bernard Williams who referred to the notion of an enlightened utilitarian elite ruling according to the principle of Sidgwick's esoteric morality as 'Government House Utilitarianism'.[173]

Now, there may well be numerous valid objections to Sidgwick's view. But that it violates the requirement that a conception of justice must be practically possible is certainly not one of them. After all, concealing the truth of utilitarianism looks like it is well within the limits of practical possibility. It is hard to see any plausible way to formulate the complaint that it is unrealistic and overly utopian. Furthermore, since the principle of utility is a maximising principle, if it were the case that concealing it from ordinary citizens turned out to be too costly, it would cease to recommend doing so. Yet the point would remain that in principle the utilitarian need not worry about satisfying the requirement of stability for the right reasons.

---

[172] *The Methods of Ethics*, 490.

[173] "A Critique of Utilitarianism," in *Utilitarianism: For and Against*, edited by J.J.C. Smart and Bernard Williams, (Cambridge: Cambridge University Press, 1973), 139.

This case points to a fundamental misalignment between the idea of stability for the right reasons and the idea of the limits of practical possibility. The conditions that a conception of justice must meet in order to pass the stability test go far beyond the conditions it must meet to be within the limits of practical possibility. In order for a conception of justice to be practically possible, there is no requirement that it must be one that could be freely and generally accepted in a WOS. In fact, the test of whether a conception of justice is practically possible is largely indifferent with respect to its content, only caring about the features that speak to its achievability. This indifference means that an appeal to the limits of practical possibility cannot explain the significance of stability. The claim that our conception ought to be practically possible cannot explain the key feature of the stability test: the requirement of free and general acceptance in a WOS.

While practical possibility may be a necessary condition for a plausible conception of political morality, it cannot be sufficient. After all, an evil dictator could ask the question of whether his nefarious plans were within the limits of practical possibility, and an affirmative answer would not imply he ought to go through with them. But satisfying the reasonable agreement principle is generally taken to also be sufficient for legitimacy. Indeed, it needs to be sufficient if it is to play the role of explaining the scope of the legitimate exercise of political power. If it is only a necessary condition, then some other idea will end up doing the majority of the work of explaining this scope.

This thought points to a possible response by Quong. A feature of his view is that he does not aim to defend a liberal conception of justice against those who doubt the plausibility of liberal views.[174] Rather, he aims to answer the question: what kind of liberalism should we accept? His arguments are thus addressed principally to those who already share liberal commitments, rather than to (for example) proponents of utilitarianism. Given this narrower argumentative target, Quong might not be concerned that the limits of practical possibility cannot be a sufficient condition for the legitimacy of the exercise of political power. After all, he might say, we have reasons to accept liberalism that are entirely independent from this idea. Those independent

[174] *Liberalism without Perfection*, 140.

reasons, together with the practical possibility test will thus be able to provide an account of the necessary and sufficient conditions for the legitimate exercise of political power. This would fit neatly with his claim, which I cited earlier, that the aim of the internal conception is to show that liberalism is coherent.

However, this response as I have stated it so far would leave the view incomplete. It seems hard to deny that there are a large number of liberal conceptions of justice that at least meet the condition of being practically possible. So we would need to know much more about our independent reasons to accept liberalism in order to make any choices within this large set.

One way to fill in these details would be to say that the independent reasons come from the argument from the original position. This gives us reason accept a certain kind of liberal theory of justice, and the stability argument just provides a test of practical possibility that is necessary but not sufficient for full justification.

This strategy would fit nicely with Rawls's general picture, whereby the justification of the conception of the justice comes in two distinct stages that are both necessary: the argument from the original position and the stability argument. I should note, however, that Quong's view is that this picture ought to be modified, such that the stability or consensus stage comes first.[175] He argues that we should begin by asking what would be the object of consensus in the WOS: on his view this is (primarily) the three political values of freedom, equality and fairness, as we saw above. Then the next step is to claim that Rawls's three liberal principles cannot be reasonably rejected by those who accept these political values. The original position, as a device that models these values, is one way to argue from the values to the principles, though there may be other ways.[176]

Regardless of whether we understand the stability component of the argument as coming at the second stage of the argument or at the first, it might appear that this reply I just suggested is available. The argument from the original position gives us reason to accept certain principles of justice and the stability argument establishes the practical possibility of these principles, such that

[175] Ibid., 161-191.
[176] Ibid., 183.

together these are necessary and jointly sufficient conditions for the full justification of a conception of justice.

The problem with this strategy, I believe, is that the original position is designed, in Rawls's thought, for the purpose of generating a conception of justice that will be stable for the right reasons. As is now well known in the literature, on some ways of designing the original position average utilitarianism is the most choice-worthy conception of justice.[177] However, there are features of Rawls's design that mean that average utilitarianism would not be chosen. Though there are a few such features, the one that most warrants mention is that the original position is not designed so that the parties choose principles of justice on the assumption that there is an equal probability of them occupying each possible social position; rather, they have no knowledge of the probabilities that they will occupy different social positions. How might this feature of the original position be justified? The bare idea of impartiality, which is sometimes thought to be the principal justification for the original position, is insufficient to justify this feature, as both ways of designing the original position are equally impartial.[178] A more promising justification, and one that coheres with much of what Rawls has to say, is that certain features of the design of the original positon are introduced in order to ensure that the resultant conception of justice is stable for the right reasons.[179] But if features of the original position are justified in this way, then the two step strategy under consideration cannot succeed. The argument from the original position is not an independent source of justificatory support for a conception of justice that can then be combined with the stability for the right reasons test, it is rather a philosophical tool purposefully designed to satisfy the stability test. In which case, we need an account of the significance of stability that shows it to be both a necessary and sufficient condition of the full justification of a conception of justice.

[177] John Harsanyi, "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *The American Political Science Review*, 69(2), 1975, 594-606.
[178] Derek Parfit, *On What Matters: Volume I*, 346-352.
[179] Joshua Cohen, "Democratic Equality," *Ethics*, 99(4), 1989, 727-751.

4.6. Normativity and Consensus

Larry Krasnoff has offered an argument which aims to tackle head on this question of why stability for the right reasons is of normative significance.[180] I will now discuss his argument in some detail. If it succeeds it could do the work of explaining how this idea of stability can be appealed to as an answer to the rationale question, and thus provide a starting point for a defence of the Rawlsian version of the reasonable agreement principle.

Stability for the right reasons represents a form of consensus; a hypothetical consensus among the idealised citizens of a WOS. The aim of Krasnoff's argument is to show that a conception of justice's availability to this form of consensus is a necessary condition of political justification.[181] His argument draws on a family of ideas about the relationship between normative theories and the beliefs of citizens, which I will now introduce.

The first of these ideas is that all normative theories aim at a form of consensus. Such theories make claims about what we ought to do and what it would be good for us to do. In claiming that some particular action is what we ought to do—say, because it is what justice requires—the theory must also hold that there ought to be a consensus around certain beliefs; namely, those that would produce this action 'under ordinary conditions'.[182] However, the theory need not hold, necessarily, that there ought to be a consensus around the belief *that the theory is true*. The example of Sidgwick's utilitarianism that was discussed above is one way of showing this. The utilitarian may not want all citizens to believe that utilitarianism is true, due to the concern they will be unable to accurately perform utilitarian calculations in their everyday life. But even if the utilitarian thinks this, they still ought to think that there should be a consensus on *some* set of beliefs; namely, whichever set of beliefs could be sufficiently easily followed by citizens, and would best approximate utility maximisation.

---

[180] "Consensus, Stability, and Normativity," *The Journal of Philosophy*, 95(6), 1998, 269-292.
[181] Ibid., 270.
[182] Ibid., 271.

Following on from this, the second idea is that of a *consensus-hindering fact*. Such a fact is one that makes our desired consensus impossible, or at least very unlikely. Since from the point of view of our normative theory this consensus must be said to be good, any consensus-hindering facts must be said to be bad. Moreover, Krasnoff tells us, a consensus-hindering fact might be *stubborn*; that is, one that we can only deal with by altering the recommendations of our theory. The interplay between normative theories and consensus-hindering facts can be seen in a number of views in political philosophy, not only in Sidgwick's defence of utilitarianism as an esoteric morality. For example, Krasnoff tells us that

> In the *Republic*, Plato recognizes that it will be extremely difficult, if not impossible, to achieve a normative consensus on the idea that only philosophers ought to rule. The ordinary citizens of Athens, Plato believes, accept the mistaken belief that they are entitled to participate in the democratic governance of the polis. To overcome this fact, Plato recommends an elaborate set of educational reforms meant to socialize the citizens of the *kallipolis* into holding the belief that only members of the highest class are entitled to rule. Education, for Plato, is an instrumental good designed not just to create philosopher rulers, but also to create a normative consensus around the idea that philosophers ought to rule. Yet even the best Platonic education cannot convince ordinary citizens that philosophers ought to rule because they are philosophers. Convincing the public that philosophy aims at real knowledge rather than idle wordplay is for Plato an impossible task: there is no way ordinary cave-dwelling citizens can or should recognize the existence of Forms as true objects of knowledge. The ordinary person's understanding of philosophy is not just a consensus-hindering fact but also a stubborn fact, one that cannot be overcome through additional education or anything else. Hence Plato recommends that his educational and political program include a number of myths and lies designed to reconcile ordinary citizens to the fact of philosophical rule. […] Ordinary citizens can be brought to believe that only members of the highest class ought to rule. But since they cannot be brought to believe that members of the highest class ought to rule because they are philosophers, they must be brought to believe that it is the gold in their veins that entitles the philosophers to rule.[183]

The stubborn consensus-hindering fact encountered by Plato's theory is based on what he takes to be a *limitation* of the reasoning capabilities of the ordinary citizens of Athens. In this way, it is similar to the consensus-hindering fact encountered by Sidgwick. Both doubt the ability of the population to hold the correct beliefs or to infer from the correct principles what ought to be done.

So far so good. The next idea Krasnoff makes use of is the possibility of a stubborn consensus-hindering fact that is unlike those considered by Plato and Sidgwick. Whereas those

[183] Ibid., 274.

facts were ascribable to the limitations of the human capacity to reason, Krasnoff broaches the possibility of a such a fact that is instead ascribable to 'the workings of reason itself'.[184] Suppose, that is, that there is a stubborn consensus-hindering fact that comes about (or remains in play) when all citizens are reasoning perfectly according to the normative theory in question. In this case the move that Plato and Sidgwick made is no longer available. When Sidgwick suggested a consensus on principles that would be simpler for ordinary citizens to follow than the principle of utility, this suggestion was justified according to utilitarianism itself. But if we were to have a stubborn consensus-hindering fact that was attributable to the *successful* exercise of our rational capacities, we could not simply recommend a consensus on a different set of beliefs instead. After all, it could not be those case that this recommendation was justified according to the theory itself, as, *ex hypothosi*, the citizens who are failing to reach consensus are reasoning impeccably. This notion is pivotal to Krasnoff's argument, so it worth ensuring it is stated precisely. In the Plato-Sidgwick case, we have two tiers of reasoners: the accurate utility calculators in 'Government House', who reason correctly according to the theory, and the general population, who fail to reason correctly. When we choose, as Sidgwick suggests, to bring about a general consensus on a simpler set of principles for utilitarian reasons, we privilege the reasoning of those in the Government House. And we do so justifiably as—from the point of view of utilitarianism—they are right and the ordinary citizens are wrong. However, if we were to have a consensus-hindering fact that was ascribable to 'the workings of reason itself', as Krasnoff is considering, then a similar move could not be justified. In this scenario, there would be no grounds for privileging the reasoning of one of the groups or parties that had failed to reached consensus, as *ex hypothesi*, they are all reasoning perfectly. An asymmetrical privileging of the reasoning of one of the groups or parties would thus be unjustifiable.

What should we do if we encounter a consensus-hindering fact of this kind? Krasnoff offers the following argument:

[184] Ibid., 275-276.

If, for a normative theory:

(1) there exists a consensus-hindering fact;

(2) the fact is stubborn;

(3) the fact is ascribable not to the limitation, but to the full exercise of individual rationality;

(4) there is not to be a conflict within reason;

Then

(C) the normative theory must be such that the beliefs it identifies as fully justified do not conflict or interfere with this consensus-hindering fact.[185]

In short, if we encounter a stubborn consensus-hindering fact that is ascribable to the proper exercise of reason, we must make our theory consistent with it.

All of this would matter very little if there were no real examples of facts of this kind for us to worry about. However, Krasnoff thinks that the fact of reasonable pluralism is exactly such a fact.[186] According to this claim, in circumstances where basic rights and liberties are protected the exercise of human reason leads to the affirmation of a variety of conflicting comprehensive doctrines. This variety or pluralism of doctrines can be explained solely by appeal to the list of burdens of judgment discussed earlier. Krasnoff argues that since the fact of reasonable pluralism satisfies the conditions in premises (1) to (3) above, and because premise (4) is true, the conclusion that the content of our normative theory must be such that it is not in conflict with range of reasonable comprehensive doctrines follows.

This argument has the potential to provide a defence of the Rawlsian reasonable agreement principle in the following way. I have been raising the question of whether an appeal to the idea of stability for the reasons might be able to provide an answer to the rationale question, by explaining the significance of agreement amongst citizens of the WOS. If Krasnoff's conclusion that our normative theory must be made consistent with the fact of reasonable pluralism amounts to the

[185] Ibid., 276.
[186] Ibid., 277-279.

claim that our normative theory must be made consistent with the variety of comprehensive doctrines that citizens of the WOS hold, then it can offer an account of the significance of agreement amongst such citizens. On this view, a conception of justice must be the object of agreement amongst such citizens because all normative theories must want there to be agreement on *some* set of beliefs, and, given that the views in the WOS have been identified as rational by premise (4), the conception of justice cannot recommend that they reject those views in favour of a consensus on something else.

Though I do not think that Krasnoff's argumentative strategy ultimately succeeds, I will not be disputing a number of the premises in his argument. In fact, his account bears some similarity to the view that I develop in the second half of this thesis. For the sake of argument, I will begin by granting premises (1) and (2). I will grant, that is, that the fact of reasonable pluralism is both consensus-hindering and stubborn. I will also register my agreement with premise (4). If a normative theory grants that this pluralism of comprehensive doctrines is ascribable to the fully adequate exercise of reason, then it would be inconsistent to hold that there ought to be a consensus on some other set of beliefs.

The claim I wish to dispute here is premise (3). Why must a normative theory hold that the pluralism that would arise and persist among citizens in the WOS is fully rational? As Krasnoff accepts, this is a very strong claim. In order to accept it we must accept, for one thing, that the various religious doctrines that would arise and persist in the WOS are the result of the successful exercise of reason. In support of it, Krasnoff states that

> given the diversity and complexity of modern societies, different persons reasoning from different points of view will, if left alone, produce a variety of comprehensive doctrines, none of which will prove universally persuasive. The inability to achieve consensus results not from a failure to reason, or from a failure of reason, but from the natural exercise of reason under conditions of pluralism. This is what is required to establish premise (3).[187]

[187] Ibid., 279.

That is, in keeping with Rawls's claims, Kransnoff thinks that the best explanation of the failure to achieve consensus in the WOS is the burdens of judgment, rather than rational failing on the part of the citizens of the WOS.

The idea a failure to reach consensus can be explained by appeal to these various ways in which our judgment can be burdened is, I think, an important part of the intuitive appeal of Rawls's view. Discussions of value and morality are so replete with disagreement that the claim that a host of these can be explained by appeal to the list of burdens is a tempting one. However, even accepting that some failures to reach consensus can be explained by an appeal to the burdens of judgment leaves open the question of whether in any particular case the sole explanation of a disagreement is these burdens, or whether there is a rational failing of some sort. Some disagreements are due to rational failings, after all.

Once it is acknowledged that this is an open question, it becomes unclear how the scope of premise (3) should be understood. It could say that a wide range of disagreements are ascribable to the full exercise of rationality, including disagreements about the correct conception of justice. But the wider the scope of the claim the fewer conceptions of justice will be compatible with the conclusion of the argument. Indeed, if the scope is very wide, then there will be no conception of justice that satisfies it. Just as there is no conception of justice that is acceptable or justifiable to all actual citizens, there is no conception of justice that is not in conflict with the views of some citizens. By contrast, premise (3) could say that there is a fairly narrow range of disagreement that is ascribable to the full exercise of rationality. Perhaps there is a single conception justice and single conception of the good that citizens would all accept if there were reasoning correctly, and the disagreement that can be explained solely by the burdens of judgment only extends to questions of how these conceptions apply to marginal cases and how they are best implemented.

Neither of these renderings of premise (3) would permit Krasnoff's argument to be used to defend the stability answer to the rationale question while consistently delivering Rawls's answer to the content question. However, the real issue is not this but rather that the ambiguity about how

the premise is to be understood renders any answer to the rationale question based on this argument inadequate. An answer to the rationale question here needs to explain why agreement among citizens of the WOS is required for the exercise of political power to be legitimate. In order to be successful it must explain why agreement among citizens of the WOS *in particular* is required, rather than agreement among agents idealised some other way. But if the range of disagreement that is ascribable to the full exercise of rationality is not specified, then this cannot be explained. A reasonable agreement principle based on agreement among citizens of the WOS will be appropriate if the range of disagreement ascribable to the full exercise of rationality is co-extensive with the range of disagreement among citizens of the WOS, but if there is any mismatch between the two then this strategy will fail. The way to respond to this objection would be to argue for a particular account of the scope of disagreement that is ascribable to the full exercise of rationality. However, since Krasnoff openly refrains from saying anything about this scope, his argument does not have any resources that could be used to argue for such an account.[188]

4.7. The Wide Reflective Equilibrium Rationale

The final account of the significance of stability that I will consider comes from Paul Weithman. He argues that showing that a conception of justice is stable for the right reasons is to show that it is in reflective equilibrium; in particular, in a reflective equilibrium that is both wide and general.[189]

As I mentioned earlier, the term 'reflective equilibrium' denotes the endpoint of a process of reasoning whereby we work back and forth from considered convictions about particular cases, the principles or rules we believe govern them, and theoretical considerations that bear on the acceptability or coherence of and between these constituent parts. The working back and forth aims not merely at coherence, but rather that these parts support and explain one another. In

[188] It is worth noting that while Krasnoff does set out and defend the argument I attribute to him in this section, his conclusion is circumspect about the possibility of this argument succeeding: Ibid., 292. His primary aim is to reply to Habermas's objection that in concerning himself with stability, Rawls had given up on the traditional project of normative theorising and moved into more pragmatic territory. For this more limited purpose I think Krasnoff's argument succeeds.

[189] *Why Political Liberalism?* 335-339.

contemporary moral and political philosophy, the method of reflective equilibrium is widely accepted as the appropriate way to reach conclusions about what principles of justice and morality we ought to accept. Indeed, T.M. Scanlon writes that

> this method, properly understood, is in fact the best way of making up one's mind about moral matters and about many other subjects. Indeed, it is the only defensible method: apparent alternatives to it are illusory.[190]

However, this terse summary of the idea obscures the two distinctions made of use of by Weithman. First, reflective equilibrium can be narrow or wide. In its narrow variant, it involves achieving the best fit between our considered moral judgments about particular cases and a set of moral principles which cohere with those judgments and from which they can be derived. In its wide variant, it involves achieving the best fit between our considered moral judgments, a set of moral principles, and also any background theories that are relevant to the plausibility of different sets of moral principles.[191] The difference between wide and narrow reflective equilibrium can be brought out as follows. Suppose that there was some general theoretical reason to think all of our considered moral judgments were false, say, because our powers of moral judgment are the result of an evolutionary process which we have no good reason to believe tracks moral truths. This fact, if it were true, would be irrelevant to the question of whether a set of moral judgments and principles was in narrow reflective equilibrium, since it is not itself a moral conviction or a moral principle. This narrow equilibrium could be achieved between a set of moral judgments and principles even if we had good reason to think all of the judgments in question were false because they are the product of a distorted process. However, this fact would be relevant to question of whether wide reflective equilibrium had been achieved. Indeed, if it were true that we had no good reason to trust our considered moral judgments, then we should discount these judgments entirely in aiming to reach wide reflective equilibrium.

---

[190] "Rawls on Justification," *The Cambridge Companion to Rawls*, edited by Samuel Freeman, (Cambridge: Cambridge University Press, 2002), 149.
[191] See John Rawls, "The Independence of Moral Theory," in his *Collected Papers*, (Cambridge, MA: Harvard University Press, 1999), 289.

The second distinction made use of by Weithman is between individual and general reflective equilibrium. Individual reflective equilibrium is reached alone, as when a single reasoner takes into account all of the arguments they are aware of for and against different conceptions of justice, and brings their considered convictions into line with a set of moral principles. General reflective equilibrium, however, requires this state of affairs to be reach by 'all actors together'.[192] Who are all the actors that must all reach reflective equilibrium together? Weithman suggests that they are the citizens of the WOS.[193] He thus suggests that establishing that a conception of justice would be stable for the right reasons amounts to it amounts to establishing that it is in a reflective equilibrium that is wide and general.[194] His argument can thus be reconstructed as follows:

(1) We ought to choose moral principles, including principles of political morality, via the method of reflective equilibrium.

(2) Reflective equilibrium should be both wide and general.

(3) When principles of political morality are stable for the right reasons, they are in wide and general reflective equilibrium.

(4) Therefore, we ought to choose the principles of political morality that would be stable for the right reasons.

I will begin by granting that we ought to choose a conception of justice via the method of reflective equilibrium. I tend to share Scanlon's view that reflective equilibrium is the only defensible method for making up our minds about many matters, and certainly matters of political morality as I stated in chapter 1. However, fully establishing this claim would require considering the various objections that have been pressed against the method in the literature, so I will instead grant it as an assumption for the time being. It seems fairly clear, further, that reflective equilibrium should be wide rather than narrow. If the equilibrium aimed for is narrow, then our considered judgments seem to be unduly protected from important sources of criticism.[195]

---

[192] Weithman, *Why Political Liberalism?* 336.
[193] Ibid., 337.
[194] Ibid.
[195] This is what Scanlon call the 'charge of conservativism' in "Rawls on Justification," 149-50.

The claims of Weithman's that I want to probe are rather that reflective equilibrium ought to be general, and that this is established by showing our principles to be stable for the right reasons. Let me begin with the alternative position of reflective equilibrium as individual. On this view reflective equilibrium is sought by a single reasoner examining her own judgments and principles and treating others only as a potential source of judgments and principles for her consideration. Would a reflective equilibrium of this kind be sufficient to fully justify a conception of justice? One consideration in favour of answering 'yes' is that once this kind of reflective equilibrium is properly understood, it is difficult to see what the alternative would be.

I should note that talk of reflective equilibrium being 'achieved' is somewhat misleading. Since the idea denotes the end point of a process of reasoning that one cannot realistically expect to know with certainty that one has reached, it is better understood as regulative ideal. We should always aim for reflective equilibrium, and settle on the set of principles and judgments that appear to be in reflective equilibrium, but we should also always stand ready to revise our judgments and principles in light of new arguments or other relevant data.

Weithman's proposed alternative to individual reflective equilibrium is equilibrium achieved by all actors together. Is this kind of unanimous convergence of reflective equilibria necessary for a conception of justice to be fully justified? Absent any further argument, this is an open question. Wide individual reflective equilibrium tells us to seek the "conception, or plurality of conceptions, that would survive the rational consideration of all feasible conceptions and all reasonable arguments for them".[196] As Scanlon notes, this is close to amounting to the trivial truth that we should choose a conception of justice by assessing every view and argument on its philosophical merits. What stops this claim from being a truism is its incompatibility with separating or leaving out any set of judgments or principles for special treatment *ab initio*.[197] I think it is this closeness to the truism that makes reflective equilibrium such an uncontroversial idea. Perhaps there is an important reason to make reflective equilibrium achieved by 'all actors together'

[196] John Rawls, "The Independence of Moral Theory," *Collected Papers*, 289.
[197] Scanlon, "Rawls on Justification," 151.

a necessary condition for the full justification of a conception of justice, as Weithman suggests, but this reason is not contained within the basic idea that is attractive to so many. The idea of reflective equilibrium alone gives us no reason to suppose, from the outset, that such unanimity is necessary. It seems that what it directs us to do instead is consider various accounts of how the reflective disagreement of others with our view should be taken into consideration.[198]

The issue with Weithman's argument, then, is that introducing the idea of wide general reflective equilibrium does nothing to explain the need for stability for the right reasons unless we have an argument for why this equilibrium must be general in his sense. Since wide general reflective equilibrium just is stability for the right reasons, on his view, the account is incomplete without a justification for the generality condition.

4.8. Conclusion

The defences of Rawls's view considered here all fail, I have argued, to offer an adequate explanation of the significance of his notion of stability for the right reasons. If proponents of the Rawlsian version of the reasonable agreement principle want to appeal to this idea in order to explain why we ought to accept the principle in this form, then they owe us precisely such an explanation. It is, after all, not obvious that the principles we act on in political life ought to be stable in this special sense, and many have denied that they must be.

The overall task of part I has been to examine defences of the reasonable agreement principle. This discussion of Rawls and his defenders concludes this task. Though I have not considered every defence of the principle, or even every aspect of it, I hope I have shown that there is still work to be done to show that any of the most prominent accounts is fully defensible. I have argued that a successful defence of the principle must answer three questions:

[198] This would perhaps involve considering the arguments made in the literature on how we should respond to disagreement among epistemic peers. See, for example: Thomas Kelly, "The Epistemic Significance of Disagreement," *Oxford Studies in Epistemology: Volume 1*, edited by Tamar Gendler Szabo and John Hawthorne, (Oxford: Oxford University Press, 2005), 167-196, and David Christensen, "Epistemology of Disagreement: The Good News," *Philosophical Review*, 166(2), 2007, 187-217.

The *idealisation question*: who are the reasonable citizens to whom the principle refers? How are they idealised?

The *rationale question*: why should we tie political legitimacy to what persons idealised in this way would accept?

The *content question*: what, if anything, would reasonable citizens accept?

None of the defences of the principle that I have examined have offered satisfactory answers to all three of these questions. Though they cannot all be said to fail for exactly the same reason. As I hope has become clear from the divergent answers to the rationale question we have considered, defenders of the principle are often defending views with quite different underlying motivations.

My critiques have focused primarily on the answers that have been given to the rationale question. We need an answer to the rationale question because we need to know why legitimacy is tied to the responses of reasonable citizens. And, moreover, given that there are a variety of possible ways of answering the idealisation question, we need an answer to the rationale question that explains why we ought to tie it to the responses of *these* idealised agents in particular. This is what each of the accounts I have considered fails to provide.

In this sense, the overall thrust of my arguments in part I concurs, to some extent, with David Enoch's recent claim that defenders of the reasonable agreement principle are unable to offer sufficient motivation for the idealisation that they employ.[199] I say 'to some extent' because Enoch aims to develop a fully general critique, applying to all versions of the requirement, while I doubt that any such general critique is possible.[200] Given the range of answers to the rationale question that have been given in the literature, there is no substitute for considering them piecemeal. As I will now argue in part II, however, I think that a compelling answer to the rationale question can be given. I begin to develop that answer in the next chapter.

---

[199] "Against Public Reason," *Oxford Studies in Political Philosophy: Volume 1*, (New York, NY: Oxford University Press, 2015), 117-130.

[200] Enoch's reason for thinking such a critique is possible depends on his view that there are very few plausible rationales for introducing idealising conditions in the first place, drawing on his more general "Why Idealize?" *Ethics* 115(4), 2005, 759-787. I set this argument aside here, as I will discuss the challenge it raises in the next chapter.

Part II

5. THE PUBLICITY REQUIREMENT DEFENDED

5.1. Introduction

As we saw in the previous chapter, Rawls's version of the reasonable agreement principle runs into difficulties when it comes to providing an account of the significance of stability for the right reasons. If such an account were forthcoming, then we would have the beginnings of a defence of the Rawlsian version of the reasonable agreement principle. This chapter aims to offer such an account. Though my argument does not focus on the stability requirement in and of itself, it defends a core component of it: the publicity requirement. A number of theories in moral and political philosophy have included a publicity requirement.[201] These requirements differ in their precise details, but they all say that moral rules or principles must be suitable for public knowledge and acceptance. Rawls's stability requirement includes a publicity requirement, as the circumstances under which principles of justice must be complied with in order to satisfy that requirement are those of public knowledge. I think the plausibility of this requirement depends ultimately on claims about our capacity to evaluate moral requirements and the circumstances under which this capacity operates free of distorting influences. I thus defend a set of such claims, which together amount to a defence of the publicity requirement.

The chapter is structured as follows. I begin, in section 5.2 by introducing the notion of a moral competence and offering an argument for the conclusion that if we were in circumstances where this competence was operating free of distortions, our moral beliefs would converge. After briefly recapping some important features of Rawls's conception of stability (section 5.3), I go on to introduce a view about the truth or correctness conditions for moral judgments which I call ecumenical constructivism (section 5.4). Over the remainder of the chapter I fill out the details of

[201] Immanuel Kant clearly made use of such an idea in his "Toward perpetual peace," *Practical Philosophy*, Translated by Mary J. Gregor, (Cambridge: Cambridge University Press, 1996), 311-352. Arguably, Hobbes did also: see Jeremy Waldron's "Hobbes and the Principle of Publicity," *Pacific Philosophical Quarterly*, 82(3-4), 2001, 447-474. More contemporary examples, other than Rawls, include Kurt Baier in *The Moral Point of View*, (Ithaca, NY: Cornell University Press, 1958) and Brad Hooker in *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*, (Oxford: Oxford University Press, 2000).

this view, and argue that it can be used to provide a compelling answer to the rationale question for the reasonable agreement principle by supporting the publicity requirement (sections 5.5-5.9).

5.2. Convergence

The first step in my argument appeals to the idea of a moral competence. Roughly, a moral competence is the ability to work out what morality requires, at least when thinking clearly and with access to the relevant information. Though what I take this competence to amount to more precisely will become clearer below, I think its existence in some form is presupposed by the widely endorsed method of reflective equilibrium. Once again, according to this method in order to make up our mind about questions of justice and morality we should work back and forth between our judgments about particular cases and the principle or principles that best explain these judgments, aiming to reach an equilibrium state where these cohere with each other at all levels of generality. If we did not have a moral competence, this method would be certain to fail.

I will assume that we do have a moral competence in some form, rather than making any argument for its existence. I want to explore instead what accepting the existence of this competence commits us to. I think that in order to sustain the belief that we have a moral competence we must also accept the following claim, the formulation of which I take from Derek Parfit.

> *Convergence:* If everyone knew all of the relevant non-normative facts, used the same normative concepts, understood and carefully reflected on the relevant arguments, and was not affected by any distorting influence, we and others would have similar normative beliefs.[202]

Why would accepting this claim be necessary to sustain the belief that we have a moral competence? If it were false, it would mean that even in these ideal conditions there would be significant normative disagreement. Given that the moral competence is the ability to work out the correct answer to the question of what justice requires, two contradictory answers to this question cannot both be correct. In order to sustain our belief in the competence in light of such disagreement we

[202] Derek Parfit, *On What Matters: Volume Two*, (New York, NY: Oxford University Press, 2011), 546.

would have to suppose that when we encountered disagreement in such circumstances only the proponents of one of the competing viewpoints was competent. However, given that all of the parties are symmetrically situated—none are influenced by any distorting factors—there are no grounds for holding that those who endorse one view possess the competence, while those who endorse another do not. Therefore, there are no reasoned grounds to suppose that anyone has a moral competence in these circumstance. We must, then, if we want to sustain our belief that we have a moral competence, hold that in these circumstances our normative beliefs would converge.

One reason that this argument might be doubted is that in claiming that we have a moral competence we need not be implying that we are guaranteed to arrive at the correct answer to the question of what justice requires, even in conditions free of distortion. The claim that we have a moral competence could simply amount to saying that there is some chance we would arrive at the correct answer in these circumstances. If this were what the idea of a moral competence amounted to then our disagreement in these conditions could be explained by chance, without rejecting the claim that each of the parties to the disagreement is competent.

The thought that disagreement in these circumstances could be explained by chance is only plausible when what is imagined is a single application of our competence to a particular question. However, I do not think *Convergence* is best understood as the claim that in a single application of the competence to a particular question our answers would converge. It should rather be understood as the claim that we would reach the same conclusions after sustained careful reflection over a period of time, which involves the repeated application of the competence to a variety of questions. Now, I think that in accepting the existence of the moral competence we should also think that sustained reflection in circumstances free of distortion leads us toward rather than away from the correct answer. It would be an odd kind of competence if its repeated application in ideal circumstances led us to being more likely to have false beliefs. If we accept this, then the claim that disagreement in circumstances free of distortion can be explained by chance will be implausible.

I think this argument commits us to either accepting *Convergence* or denying the existence of the moral competence. Since I am assuming that there is a moral competence, I will proceed on the basis that we must accept *Convergence*.

In accepting this claim, I think we should also accept the stronger claim that the beliefs we would converge on in these conditions are correct. This follows from these conditions being those in which we are not influenced by any distorting conditions, and the assumption that we have a moral competence. If we are in circumstances conducive to exercising this competence, then we should form accurate beliefs about morality and justice.

One reason this claim may be thought to be too strong is that if our powers of discernment are to some degree limited, it will not follow that the beliefs formed in conditions free of distortion will be entirely accurate: the truth about justice might extend beyond our ability to discern it. The principal reason to think this is to note that our ability to discriminate between very similar cases is limited. For example, imagine a range of acts, from A_1 to A_{1000}, where A_1 is permissible and A_{1000} is impermissible, but each individual act is indistinguishable in terms of its permissibility from both the one that precedes it and the one that follows it. It might seem that there must be a point within this range at which our action ceases to be permissible, even though no one could plausibly claim to be able to discern it. However, even in cases like this, the convictions on which we would converge in conditions free of distortion are still correct. If our powers of discernment are to some degree limited, that does not give us reason to think that any of the claims on which we would converge in circumstances free of distortion are false. It would only give us reason to think that there are some more precise truths on which we would fail to converge due to their indiscernibility. Since it will be sufficient for my argument that the beliefs about justice that would be the object of convergence in these circumstances are correct, I do not need to take a stand on this further question about whether they capture the whole truth.

I have deliberately left one thing ambiguous in the argument of this section: in discussing the claim that we have a moral competence, I have not specified the 'we' that this claim refers to.

If you employ the method of reflective equilibrium and hold that it is a way of forming correct beliefs about what justice requires, then you must at least think that *you* have this competence, and that those with sufficiently similar cognitive capacities do too. Must you also think that the competence is widely shared? Not necessarily. Maybe there is a high threshold of intelligence that must be reached in order to be competent, or perhaps a special kind of education is needed. I mean to remain agnostic about these questions at this stage.

I have argued that in order to hold that we have a moral competence we must accept *Convergence*. But is *Convergence* true? If my argument is sound then the stakes for the question may seem incredibly high: if we discover that *Convergence* is false, we will have to revise our belief that we have a moral competence in favour of a more sceptical view. However, it is worth keeping in mind that I have not yet said anything about the circumstances in which the moral competence operates free of distortion. Though discovering that *Convergence* is false may mean that we need to revise our belief in the moral competence, another possibility is that we need to revise our account of the circumstances in which the competence operates free of distortion. The fact that our beliefs would not converge under some set of conditions might indicate that these are not in fact the conditions in which our moral competence operates free of distortion.

I do not think it makes sense to try to assess the plausibility of *Convergence* prior to presenting an account of the circumstances in which our competence operates free of distortion, so I will refrain from saying anything about this question for the time being.

5.3. Stability for the Right Reasons: A Brief Recap

Getting to the point where I can defend an account of the circumstances in which our moral competence operates free of distortion will require some groundwork. The first piece of groundwork we need is a recap of Rawls's notion of stability for the right reasons. I initially introduced this idea in chapter 4, but since it will play a fundamental role in the argument of this chapter I will now briefly recapitulate its key features.

Having argued that his conception of justice would be chosen by the parties in the OP, Rawls raises the question of whether this conception would be stable for the right reasons. In order to answer this question, he appeals to the theoretical construct of a WOS: an idealised society whose institutions are effectively regulated by the conception of justice, and are recognised as just by its citizens. Every citizen of the WOS accepts the conception of justice, knows that each of their co-citizens also accepts it, and this knowledge is publicly recognised.[203] This condition is what sometimes called *common knowledge*: not only does everyone know, but everyone knows that everyone knows and so on. What is common knowledge in the WOS is not only the acceptance of the conception of justice, but also the entirety of the case for that conception. This includes both the general beliefs and modes of reasoning that support the conception, but also any arguments and theoretical apparatus that support it: in Rawls's case, this means the OP, the arguments for why it is employed and is set up as it is, and the argument from the OP to the principles of justice.[204]

For a conception of justice to be stable for the right reasons, the citizens of the WOS must choose to continue to accept it knowing everything that they know in these conditions, which we can call conditions of *full publicity*. Two key features of the citizens themselves should be emphasised. First, they have rational plans of life or doctrines. The rationality of these plans is procedural: they may include a variety of ends, but these must be arranged into a plan that meets various conditions of rationality.[205] Second, the citizens are also understood to have grown up in the idealised conditions of the WOS, going through whatever upbringing and system of education the conception of justice recommends. Given these two features, the question of whether a conception of justice would be stable thus asks whether growing up in a WOS would lead citizens to continue to affirm the conception of justice as part of their rational plan or doctrine.[206]

However, it is not enough that citizens accept the conception of justice as part of their rational plan in conditions of full publicity: they must also choose to accept it freely. If their

[203] Rawls, *Political Liberalism*, 66-71.
[204] Ibid.
[205] Ibid., 59-60; *A Theory of Justice*, 358-359.
[206] While, as I explained in chapter 4, bracketing their desire to be just for its own sake. Ibid., 499.

'acceptance' of the conception is simply a rational response to the threat of punishment, or if it is brought about by their having been indoctrinated into an ideology, then it is not free in the relevant sense.[207] How exactly this idea of free acceptance is to be specified is an issue I will return to below.

So in order for a conception of justice to be stable for the right reasons, citizens of the WOS must freely choose to accept it in conditions of full publicity. However, it is not necessarily the case that if a conception of justice is accepted in this way it is stable for the right reasons, for it may not be stable at all. For stability there must also be a motivationally effective desire to act in accordance with the conception, and this may not follow directly from the fact that it is accepted. An important example of how there could be unanimous acceptance of a conception of justice without stability is if the society faces what is known as an assurance problem. That is, suppose that each citizen's desire to act justly is conditional on her believing that a sufficient number of her co-citizens will likewise act justly. If she were to lack sufficient confidence that her co-citizens would act justly, she will not act justly in spite of her desire to do so. If this lack of confidence holds for many or all citizens, then there would not be stable compliance with the conception of justice even though each citizen accepts it.[208]

Given that free acceptance in conditions of full publicity does not necessarily imply stability, I will divide the idea of stability for the right reasons into two components. I will say that a conception of justice satisfies the *publicity requirement*, when the citizens of the WOS would freely accept it in the informational conditions I have specified. I will say that it satisfies the *stability requirement* when, as well as satisfying the publicity requirement we could also expect citizens of the WOS to continue to comply with the principles over time.

5.4. Ecumenical Constructivism

What does *Convergence* have to do with the publicity requirement? One connection simply consists in the fact that the publicity requirement is also an idea of convergence or agreement among agents

[207] Paul Weithman, *Why Political Liberalism?* 43–51.
[208] *A Theory of Justice*, 211.

in idealised conditions. The more specific connection, which I will defend in this section, is this: the circumstances in which our moral competence operates free of distortion are appropriately represented by those in which the publicity requirement is satisfied.

This is a strong claim. If we accept *Convergence*, and that the beliefs on which we would converge in circumstances free of distortion are correct, then this claim will provide a powerful argument for accepting whatever conception of justice can satisfy the publicity requirement. Moreover, if there is a conception of justice that can be shown to satisfy this requirement, that could help to dispel doubts about *Convergence* and thus about whether we have a moral competence.[209]

I will call this strong claim *Ecumenical Constructivism* (EC). Constructivist views hold that true or correct judgments about a particular domain are those that would be accepted from a suitably idealised point of view. And constructivist views differ from each other in the manner in which they understand this idealised point of view. EC holds that this is to be understood as the point of view of agents in circumstances where their moral competence is operating free of distortion.

To see what makes EC ecumenical, take this Euthyphro-style question: do reasoners in conditions free of distortion accept the correct moral principles because they are correct, or are those principles correct because reasoners in conditions free of distortion accept them? We can call the first way of answering this question, whereby the reasoners in these conditions accept the correct principles because they are correct the 'independence-answer', as it holds that the correctness of correct moral principles is entirely independent of us and our judgments, even our judgments idealised conditions. And we can call the second way of answering this question the 'dependence-answer', because it denies precisely this independence.

Constructivist views are typically taken to be committed to the dependence-answer, holding that there are no moral truths 'outside' of the relevant idealised point of view. EC, by contrast,

[209] That is, we might reason as follows: having found an account of the conditions in which this competence operates free of distortion that we find compelling, and having established that a particular conception of justice would be the object of the convergence in these circumstances, we now have warrant to increase our confidence in the claim that we have a moral competence, which we originally took as a working assumption.

explicitly rejects this commitment: it is neutral between these two ways of answering the Euthyphro question. This feature of EC forecloses some common objections to constructivist views from those who are inclined to opt for the first answer to the Euthyphro question. Even those who are strongly inclined in this way could accept that a properly designed constructivist procedure might be a powerful tool for evaluating our moral beliefs.[210] This is what makes EC ecumenical: it is compatible with these two ways of understanding the nature of moral truth.

As I say, constructivist views are not usually understood as being ecumenical in this way. In the paper that inspired much of the recent interest in constructivist views, Rawls writes:

> Kantian constructivism holds that moral objectivity is to be understood in terms of a suitably constructed point of view that all can accept. Apart from the procedure of constructing the principles of justice, there are no moral facts.[211]

Similarly, Darwall, Gibbard and Railton write that:

> The constructivist is a hypothetical proceduralist. He endorses some hypothetical procedure as determining which principles constitute valid standards of morality. […] A proceduralist, then, maintains there are no moral facts independent of the finding that a certain hypothetical procedure would have such and such an upshot.[212]

But if we understand the idealised point of view to which constructivism refers as the point of view of agents in circumstances where their moral competence is operating free of distortion, then this particular feature of constructivism can be jettisoned. The Euthyphro question can be set aside, at least temporarily, while we come up with an account of the conditions under which our moral competence operates free of distortion. If an acceptable account of these conditions can be defended, then the Euthyphro question will be without practical implications: if two parties agreed on what these conditions were, and on what we would accept if we were in them, then even if they continued to disagree about the Euthyphro question this would not in itself give rise to any disagreements about what they ought to do. They could both accept that the principles that would

[210] See, for example, Matthew Kramer, *Moral Realism as a Moral Doctrine*, (Oxford: Wiley-Blackwell, 2009), 67-68.
[211] "Kantian Constructivism in Moral Theory," *Collected Papers*, 307.
[212] "Toward *Fin de siècle* Ethics: Some Trends," *The Philosophical Review*, 101(1), 1992, 140.

be chosen in these conditions are the right ones, only disagreeing about what explains their correctness.

This reason for developing EC is strategic: if a view of this kind can be successfully spelled out, then it would allow us to provide an argument that begins with the dialectically modest premise that we possess a moral competence and takes us to the conclusion that particular principles of justice or political morality are correct. This argument would go as follows:

(1) We possess a moral competence.

(2) If we were in the circumstances in which our moral competence operates free of distortion, our normative outlooks would converge.

(3) The principles on which we would converge in these circumstances are correct.

(4) The moral competence operates free of distortion in circumstances C.

(5) In circumstances C we would converge on conception of justice J.

(6) Therefore, conception of justice J is correct.

Since this argument aims to defend a conception of justice by showing that it would be the object of convergence under particular conditions, I will call it the *convergence argument*.

The convergence argument is valid. I am assuming (1), and I have argued thus far for (2) and (3). Though the attraction of the argument will depend considerably, of course, on how (4) is specified and defended.

Before I turn to this task, I want to make one final comment on EC. When Rawls later presented his constructivism as 'political' rather than Kantian in *Political Liberalism* he jettisoned exactly the Euthyphronic commitment that I am suggesting be jettisoned: he stated, in contrast to the quoted passage above, that political constructivism is consistent with rational intuitionism as it neither asserts nor denies the rational intuitionist's claim that there is an independent order of moral values.[213] Is EC as I have characterised it therefore simply a restatement of Rawls's view?

[213] See *Political Liberalism*, 95. In spite of this, characterisations of constructivism—including those which take Rawls to be a central example of the view—have tended to specify it as committed to the dependence-answer. See, for example, the introduction to *Constructivism in Practical Philosophy*, edited by James Lenman and Yonatan Shemmer, (Oxford: Oxford University Press, 2012), 4.

The answer to this question is far from clear. EC as I have specified it has two features. The first is the ecumenical feature, and the second is the identification of the constructivist point of view with that of a citizen of the WOS in circumstances of full publicity. Though I think it is clear that Rawls accepted this first feature, it is unclear what he would say about the second. Rawls is generally understood, both by himself and by others, as a constructivist in virtue of the fact that he takes the correct principles of justice to be those that would accepted in the OP.[214] The parties to the OP, as we saw in chapter 4, are the rational representatives of the citizens of the WOS. Though the point of view of a citizen of the WOS and the point of view of the parties in the OP are related in various ways, they are also clearly distinct from each other. In the OP, the parties are tasked with choosing a conception of justice for the regulation of social cooperation in the WOS. They are understood as having a determinate conception of the good that they want to advance, though as they are behind the veil of ignorance they are unaware of the details of this conception. By contrast, citizens of the WOS are emphatically *not* behind the veil of ignorance, so they are fully aware of the details of their particular conception of the good. I think this point at least establishes that EC is not so obviously Rawls's view that it should be referred to as such. It may turn out that the best reading of Rawls takes him to accept EC, though since the aim of this thesis is not exegetical I will not take a stand on this question.

5.5. The Conditions of Undistorted Judgment

Having laid the necessary groundwork by giving a recap of the stability argument and introducing EC, I am now ready to begin defending an account of the conditions in which our moral competence operates free of distortion. Recall that the first claim I made about EC was that it holds that the circumstances in which our moral competence operates free of distortion are

[214] For Rawls referring to the original position as his constructivist device, see: "Kantian Constructivism in Moral Theory," *Collected Papers*, passim. For others doing the same, see: G.A. Cohen, *Rescuing Justice and Equality*, 277; Onora O'Neill, "Constructivism vs. Contractualism," *Ratio*, XVI(4), 2003, 319-331; Sharon Street, "What is Constructivism in Ethics and Metaethics?" *Philosophy Compass*, 5(5), 2010, 365.

appropriately represented by those in which the publicity requirement is satisfied. This is undoubtedly a controversial claim, and I will spend the remainder of this section defending it.

The publicity requirement draws together a number of conditions, so a defence of EC must show that these are each necessary and jointly are sufficient for our moral competence to be operating free of distortion. I divide these conditions into three categories: agential; informational; and background, and discuss them in turn. The agential conditions are comprised of claims about the rationality of the citizens of the WOS—claims about how they reason. The informational conditions state what information is available to them in their reasoning. And the background conditions state what circumstances they are in more widely, including claims about their upbringing and the nature of their society beyond what information is available to them. An account of the circumstances in which our moral competence operates free of distortion must be comprised of some specification of these three conditions: if we are capable of reasoning in an undistorted manner, it must be because we are the right kind of agents, with the right kind of information, in the right kind of conditions.

Agential conditions

The citizens of the WOS have plans of life or doctrines that are rational in the following way: they are plans of how best to fulfil their desires. I will therefore be arguing that in order for our moral competence to be operating free of distortion, we must have a plan of life that is rational in this way.

On some theories rationality requires us to do whatever would best fulfil our desires; that is, either our present desires or those we would have in idealised conditions. However, other theories hold that what rationality requires us to do is not given by our desires, but rather by facts about the objects of our desires. This latter view says that there are facts that make certain ends

worth pursuing, and that rationality can require us to pursue these ends even if we have no desire to do so. We can call the former desire-based view *subjectivism*, and the latter view *objectivism*.[215]

The rationality of the citizens of WOS is clearly specified by a subjectivist view. But why, it might be asked, not specify it with an objectivist view instead? It might seem that in order to defend my claim, I must also defend subjectivism as the best theory of rationality in order to avoid begging the question against the objectivist view.

I do not think this is the case, however. One consideration that militates strongly against using an objectivist view to specify the rationality of the citizens of the WOS is that doing so would make *Convergence* trivially true. If we were to say that our moral competence is distorted whenever we fail to respond appropriately to various facts about the world, this would amount to saying that our moral competence is distorted when we fail to accept true claims about what we ought to do. Saying this would make *Convergence* the trivial truth that those who accept the truth would accept the truth. But my defence of *Convergence* was not a defence of this trivial truth, it was a defence of the substantive claim that the outlooks of agents in undistorted conditions, non-trivially specified, would converge.

So if it is possible to specify the agential conditions in which our moral competence is operating free of distortion, it must be by using a subjectivist conception of rationality. However, doing this need not put the view in conflict with objectivism, let alone beg the question against it. At this point it is worth recalling the ecumenical character of EC. EC need not deny that ultimately in order to be rational a citizen must respond appropriately to the facts in the way the objectivist claims, though it need not accept this claim either. The question of which these pictures of rationality is correct is not one that it tries to answer. It only says that the subjectivist view is a component of what undistorted reasoning involves, not a complete account of rationality. This claim is not in conflict with objectivism; as Parfit notes, both the objectivist and the subjectivist can accept the following claim:

[215] Derek Parfit, *On What Matters: Volume One*, 45.

> What we have most reason to do, or decisive reasons to do, is the same as what, if we were fully informed and procedurally rational, we would choose to do.[216]

However, they would explain their acceptance of this claim in different ways. The subjectivist would hold that it is the fact that we would choose to ϕ in these circumstances that makes it the case that we ought to ϕ. Whereas the objectivist would hold the reverse: it is because we have decisive reason to ϕ that we would choose to ϕ in these conditions.[217] These two competing explanations correspond to the Euthyphro contrast I drew above when introducing EC. Since EC is neutral between these ways of answering the Euthyphro question, it is likewise neutral on this disagreement between the objectivist and the subjectivist. Therefore, in specifying the agential conditions using a subjectivist view I am not begging the question against objectivism.

Informational conditions

In conditions of full publicity all of the beliefs, arguments and modes of reasoning that support the conception of justice are common knowledge among the citizens of the WOS. I think the case for knowledge of this information being necessary for the moral competence to be operating free of distortion is strong. When we exercise our powers of judgment well, one thing we aim to do is avail ourselves of arguments and facts and modes of reasoning that are relevant to the issue at hand.

This is not to say that not knowing a relevant piece of information always distorts our judgment. There will be cases where though there is a relevant fact or argument that we are unaware of, knowing this piece of information would not have made a difference to our decision. In such cases we may not want to say that our judgment is distorted. However, our judgment is distorted by our lack of information whenever this information would make a difference to what we ultimately accept, and this thought is captured by the information available to citizens of the WOS in conditions of full publicity.

---

[216] Ibid., 62.
[217] Ibid., 62-63.

It might be objected that 'knowledge of the beliefs, arguments, and modes of reasoning *that support the conception of justice*' is too restrictive. If this arbitrarily excludes information that supports alternative conceptions of justice, then conditions of full publicity will stack the deck in favour of whatever conception of justice is being considered. But I do not think the condition should be understood in this way. Knowing the full case for a particular conception of justice requires knowing how it compares to competing conceptions. We do not know the full case for Rawls's principles of justice, for example, unless we know the case for utilitarianism too. This makes the information requirements of full publicity very high, of course. But that, I think, is exactly what we should expect from a conception of the circumstances in which our moral competence operates entirely free of distortion.

Background conditions

The background conditions of the WOS are such that the citizens choose whether or not to accept the conception of justice while free from the influence of force or punishment on their motives. The final claim I will be defending in this section is therefore that this kind of freedom is necessary for our moral competence to be operating free of distortion.

The intuition behind this claim is this: if we do have a moral competence, then we should be able to accept the correct conception of justice in conditions of freedom. That is, it should not be necessary for there to be force or the threat of punishment influencing us for us to accept it. As with the informational condition, this does not imply that the influence of force is always a distortion. Just as we can sometimes form accurate beliefs without full information, we can also sometimes form accurate beliefs in conditions where force or the threat of punishment influence our motives. So again my claim is that the presence of force or punishment is a distorting condition only if it would make a difference to what we ultimately accept.

I think that this intuitive case for including a freedom-from-force condition of some sort in our conception of the conditions in which our moral competence operates free of distortion is

strong. However, that still leaves it up for grabs how this condition ought to be specified. The publicity requirement includes a freedom-from-force condition of the following form: for the requirement to be satisfied, the citizens of the WOS must choose to maintain their desire to be a just citizen when the only forces influencing them are the educative effects of growing up in a society well-ordered by that conception of justice. If punishment, or the threat of punishment, is required to induce them to make this choice, then the requirement is not satisfied.

It might be objected that this specification is too strong. No conception of justice could command the assent of all the citizens of the WOS without making use of some system of punishment: even if all citizens wanted to act justly, they would still face what I introduced above as the assurance problem. This problem arises, recall, where all citizens want to act justly, but each individual citizen is uncertain about whether his co-citizens also want to act justly. If the desire to act justly is conditional on the (sufficient) compliance of others with principles of justice, then without confident assurance of other citizens' motives, an individual may not act justly in spite of her desire to do so. And if all citizens lack this assurance, then we might end up with the somewhat paradoxical result that in spite of all citizens wanting to act justly, none will. This problem can be solved by a system of punishment. Rawls writes:

> By enforcing a public system of penalties government removes the grounds for thinking that others are not complying with the rules. For this reason alone, a coercive sovereign is presumably always necessary, even though in a well-ordered society sanctions are not severe and may never need to be imposed. Rather, the existence of effective penal machinery serves as men's security to one another.[218]

It seems plausible to suggest that, at least in normal conditions, a system of punishment will be necessary in order to solve the assurance problem. Moreover, it seems perfectly acceptable to use a system of punishment for this purpose. Is this a problem for the freedom-from-force condition that the publicity requirement includes? I do not think so. On the system described, punishment is not required in order to induce in citizens the desire to be just. Rather, it is required solely to provide a basis for confidence in others acting justly. When a penal system is only needed to play

[218] *A Theory of Justice*, 211.

this limited role, the decision of the citizen of the WOS to accept the conception of justice is still made free from the influence force. This point can be put in terms of the distinction drawn above between the publicity requirement and the stability requirement. A penal system is not needed to ensure that a conception of justice satisfies the publicity requirement; rather, it is needed to ensure that a conception of justice that has passed that requirement also satisfies the stability requirement.

An alternative way to make the objection that this freedom from force condition is too strong would be to argue that the distinction between free and forced acceptance that it posits is a false dichotomy. What about an intermediate case, where a conception of justice is accepted due to the administration of a moderate amount of force? Why wouldn't our moral competence be operating free of distortion in such a case? Let me state this concern more precisely. Suppose that conception of justice J satisfies the freedom-from-force condition we are considering, whereas J^* would only be accepted due to the administration of a moderate amount of force. Should we rule out the possibility that J^* could be the outcome of an undistorted exercise of our moral competence? I believe that we should. Consider the question that we are imagining citizens of the WOS to ask themselves: should I maintain my desire to be a just person, even when I set aside my desire to be just for its own sake and reason on the basis of my other desires? Given that these citizens will want to avoid even moderate amounts of punishment, whether they answer this question in the affirmative would influenced by such threats. In fact, in this case the entirety of the difference in content between J and J^* is, *ex hypothesi*, explained by this influence. We can thus reverse the objector's question: why would force be necessary to induce citizens of the WOS to maintain their acceptance of J^*? It is difficult to see what explanation can be given that is consistent with maintaining that citizens of the WOS have a moral competence.

I think, then, that we should accept a freedom-from-force condition that does not allow the threat of punishment or other force to have any influence on the motives of the citizens of the WOS when they choose whether or not to maintain their acceptance of the conception of justice.

That is, we should hold this condition is a necessary component of what it takes for our moral competence to be operating free of distortion.

We might then have the opposite worry: that this freedom-from-force condition is too weak. After all, it allows that the educative effects of growing up under a particular conception of justice can legitimately influence the choice made by the citizens of the WOS. Why are these effects not also distorting influences? What distinguishes them from manipulation or indoctrination into an ideology? To answer this question, we should first note that it is inevitable that we are to some degree shaped by our upbringing and the institutions under which it takes place. If we were to hold that all educative effects were distorting, then given that some such effects are inevitable, we would have to hold that our moral judgment is inevitably distorted. This would be to hold that we do not in fact have a moral competence at all, since we are inevitably led astray. We must, therefore, accept that some effects of this kind are legitimate influences on us.

Nonetheless, it might seem that the educative effects of growing up under a particular conception of justice could amount to something like manipulation or indoctrination into an ideology, and thus ought to be taken to be distorting influences. After all, we might think that educative effects could be used to bring about the acceptance of a conception of justice the content of which we find perverse or arbitrary, and conclude that their ability to do this shows that they can be a manipulative or indoctrinating force.

To evaluate this challenge, let us suppose that we have had an upbringing in accordance with a particular conception of justice. That is, we have gone through whatever system of education and family life it recommends; or, if it takes there to be a range of permissible such systems, we have gone through one within this range. These systems may have included some coercion and deception, if the conception of justice takes such actions to be recommended or permissible components of an upbringing. Now suppose that as adults we reflect on our upbringing with full information about the arguments for this conception of justice and for why it recommended the systems it did, as per the informational condition specified above. Now, finally, suppose that with

this information, and in background conditions where neither force nor the threat of punishment have any influence on our motives, we decide that we accept this conception of justice and thus endorse whatever role it had in shaping us.

Given, as I have said, that it is inevitable that we are subject to educative effects of some kind during our upbringing, it seems to me that if we have this kind of retrospective reflective endorsement of these influences in circumstances of full information and freedom-from-force, we cannot claim to have been manipulated or indoctrinated: our capacity for rational judgment has been neither bypassed nor subverted. For this reason, then, I think we should accept that these educative effects are not themselves distorting, provided that we would later affirm them in this way.

This completes what I have to say in support of the agential, informational and background conditions that I take to constitute those in which our moral competence operates free of distortion. I said at the outset that an account of the conditions in which our moral competence operates free of distortion would have to amount to some specification of these three conditions: if we are capable of undistorted reasoning, it must be because we are the right kind of agents, with the right kind of information, in the right kind of conditions.

With the account of the three conditions before us, we can ask finally whether anything important has been omitted. One possible candidate condition is impartiality. That is, we might think that our moral competence is often distorted by our tendency to give greater weight to the interests of both ourselves and our friends, family and loved ones. And we might then take this to imply that a condition that requires us to weigh the interests of all equally in some sense ought to be included among the set of conditions in which our moral competence operates free of distortion.

In spite of any intuitive appeal this condition might have, I do not think it should be included. Different conceptions of justice take different stances on the extent to which partiality is acceptable, if at all. To include some particular degree of partiality as a distorting condition would be judge this question prematurely. It may be that only a conception of justice which includes a

strong commitment to impartiality would satisfy the publicity requirement, or it may be that permitting some partiality is necessary. What I think we should say instead is that whatever degree of impartiality would be accepted under the agential, informational and background conditions that have been specified is acceptable.

I provisionally conclude, then, both that the agential, informational and background conditions included in the publicity requirement are necessary for our moral competence to be operating free of distortion, and that there are no further conditions that are necessary. I take the argument of this section to thus constitute a provisional defence of the publicity requirement as representing the conditions in which our moral competence operates free of distortion. It is a provisional defence for a reason I have already mentioned: if we were to later discover that there is no conception of justice that could satisfy the publicity requirement, or that the content of the conception that could was unacceptable, then this would give us cause to reconsider our specification of these conditions.

Though I have provided the basic reasoning for each of the conditions, in the next two sections I will say more both about the general rationale for introducing them and about the normative force of the publicity requirement when it is understood in this way.

5.6. Challenges to Constructivism

I have argued that part of what makes EC an attractive view is its neutrality on the question of whether agents in undistorted circumstances accept the correct principles of justice because they are correct, or whether those principles are correct because agents in undistorted circumstances would accept them. Both answers to this question are to some extent controversial. Those who disagree about this question could, I have said, agree on a conception of the conditions in which our moral competence operates free of distortion. And this agreement would be significant, as they could then come to agree about what they ought to do, while continuing to disagree about the nature of the moral truth. However, constructivist views have been thought to face number of

serious challenges. In this section I will take a look at some of the most prominent of these challenges and consider what difficulties, if any, they raise for EC.

Content-based challenges

A natural worry we might have about constructivist views is about extensional adequacy. That is, we might worry that the idealised parties in the constructivist procedure will fail to accept some moral claims that we think with some certainty are true. This worry will be especially forceful if we think that the constructivist procedure will fail to vindicate what Rawls called 'fixed-points' in our moral judgment such as the wrongness of genocide and slavery.[219] Even more strongly, though in the same vein, we might be sceptical that the idealised parties in procedure would converge at all. As I argued above, this would either give us reason to think that our account of the conditions in which our moral competence operates free of distortion was mistaken, or to think that we do not in fact have a moral competence to begin with.

I want to set both of these challenges aside for the time being. In the next chapter, I will argue that the citizens of the WOS in conditions of full publicity would converge on some principles of political morality, and that they would not fail to converge on any moral fixed points. However, though I will not address either of these challenges here, they do serve to highlight the provisional nature of my conception of the conditions in which our moral competence operates free of distortion. If it turns out that either of these complaints has force, then we will have to return to the account of these conditions.

Independence-based challenges

A number of objections to constructivism target its claim that correct moral claims are correct because agents in idealised conditions accept them, rather than those agents accepting them because they are correct. Since EC is not committed to this claim, none of these objections will be

[219] *A Theory of Justice*, 17.

severely damaging to it. However, it will be valuable to briefly reflect on these objections to constructivism for two reasons. First, if the claim clearly ought to be rejected then the ecumenical feature of EC that I have been emphasising will be of no value. Second, considering these challenges will help to clarify the nature of EC.

The first of these challenges runs as follows. Whatever idealising conditions are chosen to specify the constructivist point of view, they will either be moralised or non-moralised.[220] If they are moralised, then we do not in fact have a constructivist view at all, as we have acknowledged moral constraints that are conceptually and explanatorily prior to the procedure. If they are non-moralised, then why should we expect the principles that come out of the constructivist procedure to capture our moral convictions? If the results of the procedure fail to capture our moral convictions, then the constructivist procedure will be asking us to abandon them. On this, Russ Schafer-Landau writes:

> To justify abandoning our central ethical views for those that would emerge from such a process, we would need to have greater warranted confidence in the process than in the particular views themselves. That is highly unlikely.[221]

I think Schafer-Landau is right that to moralise the idealising conditions that are used to specify the constructivist point of view would be jettison the commitment to holding that moral claims are correct because they are the outcome of that procedure. So the constructivist should avoid taking this route. However, I do not think that having more confidence in a procedure for arriving at moral judgments than we have in our moral judgments themselves is as unlikely as he supposes.

Consider the following. Suppose you have some moral judgments in which you have a great deal of confidence: say, that killing the innocent in most circumstances is wrong, that treating people differently on the basis of their race is wrong, among others. Now suppose that there are also a number of moral issues about which you are somewhat unsure, such as distributive justice and the status of non-human animals. According to the method of reflective equilibrium, we should

[220] I take this challenge from Russ Schafer-Landau, *Moral Realism: A Defense*, (Oxford: Oxford University Press, 2003), 42-43.
[221] Ibid.

work back and forth between our considered moral judgments, the principles we believe best explain them, and any theoretical considerations that bear on the acceptability or coherence of and between these constituent parts. One way this process of seeking reflective equilibrium might be represented would be to include a description of a decision procedure which implies the principles that we think best explain our considered judgments, alongside the principles and considered judgments themselves. This is exactly what Rawls's original presentation of reflective equilibrium looked like, with the OP as the decision procedure.[222] If we accept this picture as a possible route that the process of seeking reflective equilibrium could take, then we must accept that we might well be asked to give up a moral judgment that we have due to that fact that it would not be vindicated by this procedure. Could we ever justifiably have more confidence in a procedure like this than in one of our prior moral judgments? I see no reason to think not: if the procedure is a description of a decision procedure that implies principles that we find plausible, principles which in turn explain our other considered judgments, then we ought to accept that we could. To deny this would be to deny that seeking reflective equilibrium can cause us to abandon some of our considered judgments.

Let me now consider a second independence-based challenge, which is somewhat more complex. The agential, informational and background conditions I have introduced above are clearly a considerable idealisation. Suppose that instead of taking the responses of citizens in these idealised conditions to be correct, we simply took what actual non-idealised citizens accepted to be correct. This view would clearly be inadequate, for a reason I pointed out at number of points in part I: there are no moral principles that all actual citizens accept, so this view would fail to vindicate any of our considered moral judgments. However, if we are asked what our rationale is for idealising the citizens, we cannot appeal solely to the fact that taking the acceptance of actual citizens to be dispositive would make our view inadequate in this way. Doing this would make the idealisation simply an ad hoc fix, introduced solely for the purpose of achieving extensional

[222] *A Theory of Justice*, 15-19.

adequacy with respect to our considered judgments.[223] Therefore, if we are going to idealise the citizens, as we must, we need a rationale for doing so that goes beyond this.

What might that rationale be? A natural rationale for introducing idealising conditions is that they are needed in order to make an epistemic procedure reliably track the facts.[224] However, if we want to hold that the acceptance of the idealised agents is not merely a guide to a domain of facts that is independent of their judgment, but rather that what it is to be correct is—constitutively—to be accepted by idealised agents, then we cannot appeal to this rationale, for it appears to presuppose that the truth is independent of our judgments. Therefore, what is needed is a rationale for idealising that explains why the idealising conditions are introduced in a non-ad hoc manner, without invoking the idea that we are trying to accurately perceive some moral facts that are independent of us and our responses. This challenge is made especially forceful by the fact that it looks like the introduction of each of the idealising conditions that I specified above could be explained by saying that they are the conditions that need to be in place for us to accurately discern an independent truth.

On this challenge, I am tempted to simply say that the rationale for introducing the various idealising conditions is that they are plausible candidates for conditions in which our moral competence operates free of distortion. If this presupposes the independence-answer to the Euthyphro question, then this is not damaging for my central thesis, as EC is compatible with either answer. Though, of course, if giving the dependence-answer was clearly mistaken for the reason under discussion, then the ecumenical feature of EC would be somewhat pointless.

Does EC presuppose the independence-answer in this way? Although I do not have a firm answer to this question, I think there is some reason to doubt that it does. The right way to make up our mind about moral questions is, I have suggested, the method of reflective equilibrium. In seeking the best fit between the principles we affirm and our considered moral judgments, it seems we are required to do a number of things. One is to avail ourselves of information that is relevant

[223] This challenge comes from David Enoch, "Why Idealize?" *Ethics*, 115(4), 2005, 759-787.
[224] Ibid., 761.

to our judgment about which principles to accept or which judgments to make. Another is to ensure we are following good norms of procedural reasoning. And another is to ensure that that our principles and convictions are not unduly influenced by contingent features of our circumstances. If that is right, then the idealising conditions I have introduced can be motivated by the fact that they offer the best explanation and explicit statement of what reflective equilibrium requires of us. Would *this* rationale presuppose that the independence-answer to the Euthyphro question? It is at least not obvious to me that it does,[225] and so I think the ecumenical nature of EC retains some utility.

5.7. The Normative Force of the Requirement

I now want to say a little more than I have done so far about the normative force of the publicity requirement when it is defended in the way that I have suggested.

The appeal to the possibility that our powers of judgment could be unduly influenced by various distorting features of our circumstances represents an attempt to utilise the force of what have come to be known in the philosophical literature as genealogical debunking arguments. These arguments appeal to the origins of our evaluative beliefs in order to undermine their justification. The basic idea here is best explained by considering the most discussed form of this argument, which appeals to evolutionary origins of our moral beliefs in an attempt to undermine their justification.[226] The evolutionary debunking argument begins by noting that our evaluative standards, including our moral beliefs, have had their content shaped by evolutionary forces. The standards we hold are the product of an evolutionary process, and were therefore developed to enhance our prospects of surviving and reproducing. What is troubling, according to proponents of such arguments, is that had a different set of evaluative standards better enhanced our prospects

[225] Ultimately this will hang on whether the commitments of reflective equilibrium can be adequately stated without the claim that the truth is independent of us. For some scepticism, see Ibid., 774-775.
[226] The literature on these arguments is now significant. See Richard Joyce, *The Evolution of Morality*, (Cambridge, MA: MIT Press, 2006), and Sharon Street, "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies*, 127(1), 2006, 109-166, for two of the most important elaborations of it.

of surviving and reproducing, we would have had different evaluative beliefs, including moral beliefs. Unless we can provide an explanation as to why it would be the case that natural selection tends to favour beings who have the evaluative capacities to reach true conclusions about morality, we lack grounds to suppose that we have such capacities. In the absence of such an explanation, then, the proponent of the evolutionary debunking argument holds that we lack sufficient grounds for our moral beliefs.

One issue with evolutionary debunking arguments is that, if successful, they seem to debunk all of our evaluative beliefs. It is not just our moral beliefs, after all, that reflect standards that are the product of natural selection. They therefore appear to licence a more general scepticism if sound.[227] Since one of the starting assumptions of my argument here has been that we have a moral competence, understood as a capacity to form correct moral beliefs, I will not focus too much attention on whether there is a good reply to evolutionary debunking arguments in general. However, one promising line of defence has been to argue that the capacity to reason *in general* would be favoured by natural selection, as being able to form true beliefs about some domains would help us to survive and reproduce. And once we have the capacity to reason in general, we are capable of forming true beliefs about other domains, such as morality, even if the capacity to form true moral beliefs in particular would not itself be favoured by natural selection. In this way, the moral competence would piggyback on the evolution of a more general rational capacity.

Supposing that it is possible to prevent the evolutionary version of the genealogical debunking argument from licensing a thoroughgoing scepticism, it has been argued that evolutionary considerations might undermine some of our moral beliefs but not others. That there are local cases of this phenomenon seems fairly clear. Many people have a very strong negative intuitive reaction to incestuous sexual relationships. And there is a simple evolutionary explanation for how strong and widely held this response is: for most of our evolutionary history, sexual relationships between close relatives would have been liable to produce offspring that had

[227] See the discussion in Katia Vavova, "Debunking Evolutionary Debunking," *Oxford Studies in Metaethics: Volume 9*, 2014, 76-101, and Guy Kahane, "Evolutionary Debunking Arguments," *Nous*, 45(1), 2011, 103-125.

decreased chances of survival. However, when we are forming a judgment about a case in which an incestuous sexual relationship is consensually embarked on in circumstances where reproduction is not possible and no harm comes to any party, it seems that we should discount our intuitive reaction, if we have it, that this is wrongful.

The point of this detour into evolutionary debunking arguments is this. That evolutionary considerations might undermine some of our moral judgments is simply a component of the wider phenomenon that genealogical considerations in general have the potential to undermine our moral judgments. This wider phenomenon has been a part of moral and political philosophy since at least the attempts to answer Thrasymachus in *The Republic*.[228] And moreover, it is a familiar part of the practice of moral reflection that we ask ourselves to what extent our moral convictions are the product of arbitrary features of our background or upbringing, and reject those which do not hold up to this critical reflection.[229] The normative force of the publicity requirement consists in it offering a potential reply to these challenges. By setting out a conception of the circumstances in which our moral competence operates free of distortions, it aims to tell us which our convictions can survive these and other genealogical critiques.[230]

5.8. Back to the Reasonable Agreement Principle

Having offered my defence of the publicity requirement by appeal to ecumenical constructivism, I am now in a position to reconsider the reasonable agreement principle that was the object of discussion in part I. Recall that this principle states that the exercise of political power is legitimate only when it is exercised in a manner that all reasonable citizens may be expected to accept. In part I, I argued that a defence of this principle would have to answer three questions. The first of these is the idealisation question. It must tell us who the reasonable citizens the requirement refers to

[228] See Plato, *The Republic*, translated by G.M.A. Grube, revised by C.D.C. Reeve, (Indianapolis, IN: Hackett, 1992), book I.
[229] Roger Crisp, *Reasons and the Good*, (Oxford: Oxford University Press, 2006), 89.
[230] In this way, the argument here has an analogous structure to an argument made by Peter Singer and Katarzyna de Lazari-Radek, which aims to show that genealogical considerations undermine every view other than utilitarianism. See their "The Objectivity of Ethics and the Unity of Practical Reason," *Ethics*, 123(1), 2012, 9-31.

are: in what ways are they idealised? The second question is the rationale question: why have we idealised them in this way? What are the grounds for tying the answer to a normative question about political legitimacy to acceptance under these particular idealised conditions? The third question was the content question: what conception of justice could reasonable citizens be expected to accept, if any?

Rawls and his defenders, whose views I considered in chapter 4, want to answer these questions as follows. On the first question, they want to hold that the reasonable citizens to whom the reasonable agreement principle refers are the idealised citizens of the WOS. On the second question, they want to hold that the reason we must make the exercise of political power acceptable to this consistency is because this is needed to show that our conception of justice is stable for the right reasons. On the third question, they want to hold that citizens idealised in this way would accept a conception of justice that included some specification of the following three liberal principles: (1) ascribing to all citizens' familiar individual rights and liberties, such as rights to freedom of expression, liberty of conscience, and free choice of occupation; (2) giving special priority to these rights and liberties, especially over demands to further the general good or perfectionist values; (3) assuring for all citizens' sufficient all-purpose means to make effective use of their freedoms.[231]

However, as I argued there, stability for the right reasons is not a satisfactory answer to the rationale question in and of itself. What we learn when we learn that a conception of justice is stable for the right reasons is pretty much just that it would be accepted by the idealised citizens of the WOS. Since the rationale question asks why we ought to care about acceptability to the idealised citizens of the WOS, the idea of stability for the right reasons alone is not fit to provide an adequate answer to it: it is simply a restatement of the claim in question. Further, as I also argued, the various explanations that have been offered in the literature as to why we ought to care about stability for the right reasons are also inadequate.

[231] *Political Liberalism*, 6.

The defence of the publicity requirement that I have offered in this chapter can, I believe, offer a satisfactory answer to the rationale question and therefore take us some way toward a defence of the Rawlsian reasonable agreement principle. To show this, let me now consider how the argument of this chapter can respond to the various issues I identified with the appeal to stability in chapter 4.

Alongside the problem that stability for the right reasons alone is unfit to play the role of answering the rationale question, there were three further issues with the appeal to stability that I considered. The first was that we might think, with G.A. Cohen, that the question of whether or not a conception of justice is capable of being stable in the WOS is irrelevant to the question of whether it is genuinely *justice*.[232] Justice and stability are simply two different things. The second was that we might think that the best moral conception *by its very nature* would be unable to order a society in this way, as public knowledge of it would undermine its ability to achieve its aims, as Sidgwick thought of his utilitarianism.[233] The third was that we might think that stability for the right reasons is a desideratum, but it is one that must be weighed against others in the process of working out which is the best justified conception of justice, as Raz claims.[234] I will now reconsider these in turn.

In order to be stable for the right reasons, a conception of justice must satisfy the publicity requirement. The defence I have offered of the publicity requirement solves the first issue identified with the appeal to stability for the right reasons as a way of answering the rationale question. Rather than trying to explain why we ought to care about what the idealised citizens of the WOS would accept by appealing to a property that is essentially a restatement of claim that the idealised citizens of the WOS would accept it, my defence of the publicity requirement appeals to the deeper idea of EC. If we accept that we have a moral competence, that our moral outlooks would converge in circumstances free of distortion, and that what we would converge on in these circumstances is

[232] *Rescuing Justice and Equality*, 327-31
[233] *The Methods of Ethics*, 489.
[234] Joseph Raz, "Facing Diversity: The Case of Epistemic Abstinence," 16.

correct, then we have grounds for accepting the basic idea of EC: that we ought to accept the conception of justice that would be the object of convergence in the circumstances where our moral competence operates free of distortion. These are premises (1), (2) and (3) of what I have called the convergence argument. If, as well as accepting the basic idea of EC, we accept that the publicity requirement appropriately specifies the conditions in which our moral competence operates free of distortion —premise (4) of the convergence argument—then we will have good grounds for accepting the conception of justice that satisfies the publicity requirement, if any do. This appeal to EC can explain, then, in a way that a naked appeal to stability for the right reasons could not, why we ought to tie legitimacy to what would be accepted by the citizens of the WOS.

Moving on to Cohen's concern, we can say that although justice and stability might initially seem to be unrelated, the convergence argument shows that they are not. Though we might begin thinking about justice with some firm judgments, there will also be various cases that lead us to be unsure which principles of justice are the right ones. We ought to respond by considering what conception of justice we would accept if we were in circumstances where our moral competence was operating free of distortion. If we accept that these circumstances are appropriately represented by the publicity requirement, then we must accept that the truth about justice and the publicity component of the stability requirement are not unrelated. Of course, if we accept the independence-answer to the Euthyphro question, then we will not think that the correct conception of justice is identical with the conception that would satisfy the publicity requirement, but we must at least think that working out what could satisfy the publicity requirement is how we ought to work out what conception of justice to accept, understood as entirely independent from us and our judgments metaphysically.

Next, to Sidgwick's claim that we might well have to hide the correct moral principles from ordinary citizens because this is what those principles, adjusted to facts about their ability to follow those principles, recommend we do. It might appear that the argument I have offered gives us grounds to reject this claim directly. If our moral conception is self-effacing in this way, if full

knowledge of it would undermine its ability to achieve its aims, then it cannot satisfy the publicity requirement and thus ought to be rejected. However, I think the situation is actually somewhat more complicated than this. The reason for this refers back to a claim I made earlier about the scope of the 'we' who possess a moral competence. Though I left this unspecified above, I can now say more about it. If we accept that the publicity requirement appropriately represents the circumstances in which our moral competence operates free of distortion, then it seems we have to think that citizens of the WOS have the competence. But I think it would be a mistake to understand this as meaning that only these idealised citizens have this competence. It is more accurate to say instead that the point of view of a citizen of the WOS is one that we, ordinary as we are, can take up in in our reasoning about what we ought to do. Strictly speaking then, if we accept that the publicity requirement accurately represents the conditions in which our moral competence operates free of distortion, we ought to think that this competence is possessed by everyone who is capable of taking up this point of view and reasoning from it accurately.

If Sidgwick's claim is understood as the claim that the correct moral principles must be kept hidden from citizens even in the idealised conditions of the WOS, then my defence of the publicity requirement implies that this claim is false. In such conditions, I have argued, our moral competence operates free of distortion and thus the correct moral conception would be accepted. However, if the claim is to be understood as holding that deception is necessary only in conditions that are less idealised than those of the WOS, then my defence of the publicity requirement alone does not provide a reason to reject this claim. Whether deception is permissible or even required outside of these idealised conditions is an open question, the answer to which depends on the content of the conception of justice that would be object of convergence. In principle, that conception could permit or require widespread deception.[235]

[235] In the same vein, Brad Hooker writes: 'Defining ethics in terms of principles or rules for public acceptance does *not* eliminate the possibility of permissible secrecy. Publicly accepted rules might be such as to authorize doing things in secret in some contexts, and even authorize lying to protect secrecy in some contexts.' "Publicity in Morality: A Reply to Katarzyna de Lazari-Radek and Peter Singer," *Ratio*, XXIII(1), 2010, 115. Emphasis in original.

Finally, to Raz's claim that stability for the right reasons is at best one value among others. This defence of the publicity requirement shows, I think, that it should not just be treated in this way. If a conception of justice would not by accepted in the conditions in which our moral competence operates free of distortion, this provides us with a strong reason to reject it rather than with the kind of loss of plausibility that might be compensated for by some other virtue.

5.9. Conclusion

In this chapter I have defended the publicity requirement by appeal to ecumenical constructivism. This view has three features. First, it says that we ought to accept the conception of justice that we would converge on if we were in the circumstances in which our moral competence operates free of distortion. Second, it says that these circumstances are appropriately represented by the publicity requirement. Third, it takes no stand on the Euthyphro question of whether the correctness of this conception of justice is independent of our idealised judgments or not.

Since the publicity requirement is a version of the reasonable agreement principle, asking what a consistency of idealised agents would accept, my defence of it amounts to a provisional defence of that principle. This defence is only provisional because, as I have noted, we may come to doubt the conception of the conditions in which our moral competence operates free of distortion that I have defended if we doubt that there would be convergence on moral claims that we think with some confidence are true. In chapter 6 I aim to address this concern by considering what conception of justice could satisfy the requirement.

6. CONVERGENCE, FALLIBILITY, SATISFACTION

6.1. Introduction

An important attraction of my defence of the publicity requirement in the previous chapter, I noted, was that it allows us to make the convergence argument, which runs as follows:

(1) We possess a moral competence.

(2) If we were in the circumstances in which our moral competence operates free of distortion, our normative outlooks would converge.

(3) The principles on which we would converge in these circumstances are correct.

(4) The moral competence operates free of distortion in circumstances C.

(5) In circumstances C we would converge on conception of justice J.

(6) Therefore, conception of justice J is correct.

The argument of the previous chapter has offered a defence of premises (2) and (3), and a provisional specification and defence of the circumstances referred to in premise (4). The purpose of this chapter will be to examine premise (5). What conception of justice, if any, would citizens of the WOS converge on in circumstances of full publicity?

As I noted in the previous chapter, one reason we might doubt the publicity requirement is that we doubt whether there is anything that citizens idealised in this way would converge on. Or, alternatively, we might think that they would converge on some principles, but these principles would fail to vindicate moral claims that we think with some certainty are true. I think that either of these results would give us reason to doubt that the publicity requirement accurately represents the circumstances in which our moral competence operates free of distortion. The only way to respond to these possibilities is to demonstrate that some principles of justice would be converged on, and that they would not be clearly inadequate in this way. That is the aim of this chapter.

I begin, in section 6.2, by stating more precisely what an argument that established convergence would need to look like. Section 6.3 then goes through the steps of Rawls's Kantian

congruence argument. This argument, whatever its weaknesses, has exactly the right structure to establish the kind of convergence we are looking for—so pinpointing where it goes wrong, which I aim to do in section 6.4, is in an important first step in trying to establish whether any similar argument could succeed. In section 6.5 I offer my own structurally analogous argument. Section 6.6 concludes by arguing that this argument leaves us with an acceptable answer to the content question.

6.2. Convergence and Desires

In order for the publicity requirement to be satisfied, the desires of the citizens of the WOS must converge. I begin by noting that convergence in the object of desires can be thicker or thinner.[236] Convergence is thicker when the shared desires are specified more specifically, and thinner when they are specified more generally. When our desires diverge at one level of generality they can still converge at another. If Alf desires strawberries and Betty desires apples, they still converge in desiring fruit generally, when the kind is unspecified. The convergence in cases like this may take different forms. On one form, Alf might agree that if he were in Betty's position, with her tastes, he too would desire apples. In such a case what they agree on is the desire to follow some procedural rule to settle the issue in question, which in this case is perhaps simply the desire to be procedurally rational: a desire to do whatever will best satisfy their more particular desires. Alternatively, Alf might hold that he wants to be the kind of person who desires strawberries even if his tastes are such that he prefers apples. This may seem somewhat odd when applied to the case of tastes, but it is less odd when applied to the moral case. It is, for example, quite plausible to want yourself not to be cruel even in the hypothetical scenario where you want to be.[237]

An argument that aims to establish the convergence of desires might therefore proceed by establishing that citizens of the WOS would converge in having a general desire with procedural content, as when Alf and Betty both want to use the same principle to decide what kind of fruit to

[236] This point is taken from David Sobel, "Do the desires of rational agents converge?" *Analysis* 59(3), 1999, 140.
[237] Allan Gibbard discusses desires of this form in *Wise Choices, Apt Feelings: A Theory of Normative Judgment*, 167.

buy. Their agreement in this case is simply agreement on the desire to be instrumentally rational. This particular desire contains a decision rule that denotes what is and is not relevant to the settling of the question of what one ought to desire, and holds that their tastes are decisive in this case. Other procedural desires will contain different decision rules, which will denote different criteria as relevant for the settling of different questions. If it could be shown that citizens of the WOS would converge on a procedural desire that included a rule with criteria for settling questions of how they ought to live their lives, then we will have taken a step toward showing that their desires would converge. Of course this discussion does not give us any reason to suppose that citizens of the WOS would converge on a desire of this form, but what it does do is allow us to see a route that an argument for convergence could take.

6.3. The Kantian congruence argument

A.

Part III of Rawls's *A Theory of Justice* attempts to establish that justice as fairness satisfies the stability requirement, and thus the publicity requirement. It does so by arguing for two conclusions: first, that citizens who grew up in a society well ordered by justice as fairness would develop a sense of justice informed by it; second, that when they bracket their desire to be just for its own sake, they will still judge it rational to maintain this desire to be just as highest-order regulative desire. I have yet to explain the notion of a highest-order desire mentioned in this second step. We can have higher-order desires in that we can desire to desire things; e.g., we can wish that we desired ice cream less than we actually do.[238] It is possible to have numerous higher-order desires which conflict both with each other and with our first-order desires. Desires vary in strength, and like first-order desires higher-order desires may ultimately be trumped by other desires in the process of deciding how to act. A high*est*-order desire is like a higher-order desire in that it is a desire about our other desires. However, it is distinct in that we can only satisfy it by treating it as regulative of

[238] This account draws on Harry Frankfurt's "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy*, 68(1), 1971, 5-20.

all of our desires. We cannot satisfy such a desire by weighing it against our first and higher-order desires, seeing what is strongest on balance. We can only satisfy it by treating it as supremely regulative. Consider, for example, the desire to live a rational life. We could not claim to be satisfying this desire if we treated it as but one of our ends to be weighed against others, the very nature of the desire, what it is a desire for, means that to satisfy it we must treat it as regulative.

The principal argument that aims to establish the citizens of the WOS would have a highest-order desire to live according to the principles of justice as fairness has been called the Kantian congruence argument (KCA hereafter).[239] This is referred to as a congruence argument because it aims to show that the principles of justice as fairness are congruent with the good of the citizens of the WOS. Though it is not the only argument Rawls makes for the conclusion that his principles of justice satisfy the stability requirement, it is the most important. If it is successful, it shows that citizens who grew up in a society well-ordered by the principles would have a desire that is *practically identical* with the desire to be a just person, and that this is a desire that they can only satisfy by accepting the principles of justice and treating them as regulative of their other desires.

The first part of this chapter will be dedicated to discussing the KCA in some detail, with the aim of understanding precisely where it goes wrong, and of seeing whether any similar argument could succeed. However, this focus on the KCA may seem somewhat odd to those who are familiar with Rawls's work. Rawls thought that the KCA failed, and offered an alternative account of stability in *Political Liberalism*, which was based centrally on the notion that a conception of justice could be stabilised by being the object of an overlapping consensus of the doctrines held by citizens of the WOS.[240] Why not focus attention instead on this argument? The answer is that I do not think that Rawls's comments on the idea of an overlapping consensus constitute a fully specified argument for the conclusion that a particular conception of justice, or one conception within a

[239] This term for the argument originated in Samuel Freeman, "Congruence and the Good of Justice," *The Cambridge Companion to Rawls*, 277-316. However, in my discussion of this argument I will primarily follow the account of it in Paul Weithman, *Why Political Liberalism?* 183–233.
[240] 133-172.

certain range, would satisfy the stability requirement in the way that the KCA does. In order to fully state such an argument using Rawls's ideas in *Political Liberalism*, I think that it would be necessary to import at least the form of the congruence argument from *A Theory of Justice*. Therefore, I think that the KCA is an appropriate focus for our attention.[241]

In order to set out and evaluate the KCA, the first thing that needs to be done is to go into more detail than I have so far about the OP. As we have seen, the OP is a deliberative procedure in which the parties are tasked with choosing principles of justice to publicly order a WOS. The principles that they choose are to be in the final court of appeal for settling competing claims in the WOS in perpetuity. That is, no higher standards than those principles are to be acknowledged at any time. Recall that these parties are also the rational representatives of the citizens of the WOS, so they seek to choose principles that will further those citizen's interests as best they can. However, being situated behind the veil of ignorance, they do not know the race, ethnicity, gender, age, income, wealth, natural endowments or conception of the good of the citizen that they represent. All they know are general facts about human social life, including the uncontroversial conclusions of science, economics and psychology, and that the citizen that they represent is instrumentally rational and possesses a determinate conception of the good.[242]

Tasked with choosing in this way, the parties are said to reason as follows. Even though they do not know the particular conception of the good of the citizen that they represent, knowing that they are rational and that they do have some determinate ends, they know they will have an interest in carrying out their rational plan whatever it happens to be. Regardless of what our particular ends are there are certain goods that it is rational for us to want; those goods the possession of which makes the successful execution of our plans more likely: liberties, opportunities, income, wealth, and self-respect. Rawls refers to these as primary goods.[243] The

---

[241] I think that Paul Weithman, who has commented more extensively than anyone else on Rawls's stability arguments, would concur with the sentiment of this paragraph. See *Why Political Liberalism?* which takes a similar approach throughout, and especially the comments on 273-274.

[242] For Rawls's most up to date description of the OP, including the perpetuity and final court of appeal conditions, see *Justice as Fairness: A Restatement*, 80-89.

[243] Ibid., 57-58.

parties to the OP therefore choose principles of justice for the distribution of these goods (or the means to these goods) with the interests of the citizen of the WOS they represent in mind, despite not knowing their particular conception of the good.

Now, as we have seen, Rawls argues that justice as fairness would be chosen by the parties to the OP over a set of other potential principles. Since he thinks this, one way to describe the desire to follow Rawls's principles is as a desire to follow the principles that would be chosen in the OP. Further, since Rawls thinks that the principles that would be chosen in the OP are principles of justice the following three desires come to one and the same thing for him: the desire to be a just citizen; the desire to comply with the principles of justice as fairness; and the desire to follow the principles that would be chosen in the OP.[244] These desires all have the same content, if Rawls is right that his two principles would be chosen in the OP, as what they all amount to is a desire to follow those principles.

That the content of these desires lines up in this way is fundamental for the KCA, as its strategy is to argue that the citizens of the WOS would have another desire, the content of which also lines up with these three. What is this desire? Rawls calls it the desire to express our nature as a free and equal rational being, but for brevity I will refer to it as the desire to express our rational nature.

Seeing yourself as having a rational nature means seeing yourself as not being inevitably bound to, or identical with any particular interest or set of interests you might have: it means seeing yourself instead as a being who can revise and alter her ends. Those who see themselves in this way take themselves to be able to exercise rational authority over their desires, organizing them into a single scheme of conduct by weeding some desires out and prioritising others.[245] And they also see themselves as capable of limiting the pursuit of their interests in line with principles of justice. In sum, seeing yourself as having a rational nature involves seeing yourself as capable of

[244] The claim that the desire to comply with principles of justice can be given a diversity of descriptions is emphasised by Weithman, *Why Political Liberalism?* 118 and 125 (for example).
[245] Here I follow the description in Robert Taylor, "Rawls's Defense of the Priority of Liberty: A Kantian Reconstruction," *Philosophy & Public Affairs*, 31(3), 2003, 259-60.

choosing your ends as an exercise of rationality, within the limits set by a conception of justice. How do you *express* your rational nature? Rawls writes:

> To express one's nature as being of a particular kind is to act on the principles that would be chosen if this nature were the decisive determining element.[246]

Therefore, we express our rational nature when we act on the principles that we would chose if our rational nature was the decisive factor determining the choice.

At this stage it is probably not especially clear what, if anything, the desire to express our rational nature has to do with the three desires I mentioned above: the desire to comply with the principles of justice as fairness, the desire to be a just person, and the desire to follow the principles that would be chosen in the OP. As I said, Rawls thinks that the desire to express our rational nature also has the same content as these three desires. But why would that be true? This is because the OP is set up such that our rational nature is the decisive determining element of the choice.

The parties in the OP choose principles of justice from behind the veil of ignorance. They do not know the particular ends or interests of the citizen of the WOS that they represent, which means that they conceive of that citizen instead as having an interest in carrying out their rational plan whatever it happens to be. Seeing yourself as having a rational nature involves seeing yourself as capable of revising and altering your ends, rather than as wedded to some particular ends. Therefore, the parties to the OP choose principles solely on the basis of the rational nature of the citizen that they represent: they have, by stipulation, no other information to go on. It is primarily the veil of ignorance, then, that makes the OP a scenario in which our rational nature is decisive in determining the choice. [247]

The veil of ignorance is not the only condition that makes it the case that the OP is a choice scenario in which our nature as persons is the decisive determining element. What completes the argument for the claim that citizens of the WOS's desire to express their rational nature is practically identical with the desire to act according to the principles that would be chosen in the

[246] *A Theory of Justice*, 222.
[247] Ibid., 223; Weithman, *Why Political Liberalism?* 222-223.

OP is the condition I noted above: that principles chosen in the OP are to be the final court of appeal for the settling of competing claims in the WOS, in perpetuity. Why would the citizens of the WOS think that acknowledging principles that are the final court of appeal for settling their competing claims in perpetuity was a part of their nature? This is, at the very least, far less clear than the link between the veil and the desire to express their rational nature.

What must first be noted is that framing a rational plan of life in accordance with principles that are taken to be the final court of appeal in perpetuity is to acknowledge those principles as supremely regulative, as in the notion of a highest-order desire described above. This is because when we acknowledge certain standards as the final court of appeal in our reasoning there are, by stipulation, no higher standards against which to check that reasoning. However, acknowledging these standards as the final court of appeal need not imply that we acknowledge them as such over our complete life, as we might acknowledge different highest-order standards at T_1 and T_2. But if we acknowledge principles as the final court of appeal *in perpetuity*, we acknowledge them as supremely regulative over our complete life.

To see why acknowledging principles that occupy such a place in our rational plan answers to citizens' of the WOS's nature as persons, we must first know what the alternative to acknowledging such principles is. When citizens of the WOS acknowledge principles of justice as supremely regulative, they draw up a rational plan of life within the scope that those principles allow. When the principles of justice say that certain ends are unjust, those ends are ruled out as potential candidates for inclusion in the plan. Suppose that instead they draw up their plan without acknowledging any principles of justice as supremely regulative. When drawing up a plan in this way, *any* ends would, at least potentially, be viable candidates for inclusion, as none would be ruled out by regulative principles of justice.[248]

The contrast between these two ways of drawing up a rational plan of life corresponds roughly to the contrast between teleological and deontological views in ethics. The former put the

[248] *Why Political Liberalism?* 212.

good prior to the right, defining moral duties and obligations in terms of the good as in the paradigmatic case of utilitarianism. The latter take the right to be prior to the good, as in the case of justice as fairness where moral duties and obligations are specified independently. When we apply a deontological view and draw up a rational plan of life in accordance with it, we exclude any ends that conflict with its principles of justice. This means that the rational plans of citizens are not in conflict from the point of view a deontological view, for Alf cannot have a claim to satisfy his end that conflicts with Betty's claim to satisfy her end, and so on. On teleological views, by contrast, such conflicts can arise: no ends are ruled out in advance by principles of justice, so any ends are potential candidates for inclusion in a rational plan, including ends that conflict with the claims of others. Teleological views have ways of settling these conflicts, of course, such as by appeal to the principle of utility.

To further illustrate this contrast, imagine two well-ordered societies. In the first, citizens acknowledge principles of justice as regulative of their plans, and so their competing claims can be settled by reference to a regulative conception of justice that they each acknowledge. If Alf has an end within his plan that precludes the achievement of an end that Betty has as part of her plan, then Alf's end is either consistent with the principles of justice or it is not. If it is not, then it is an end that Alf should not have and his claim to its satisfaction is illegitimate. In the second society, there are no such acknowledged regulative standards to settle competing claims. Conflicts are therefore settled by appeal to a teleological principle of right such as the principle of utility. Rawls argues that part of the citizens of the WOS's desire to express their rational nature is a desire to live in the former kind of society. And the explanation of this desire is that rational citizens of the WOS have an interest in *the unity of their plan of life*, and this unity can only be achieved where principles of justice are treated as supremely regulative. He claims that unified plans have the following four features.

(F1) They are dictated by rational choice, rather than by purely preferential choice.

(F2) The ends included are consistent with our judgments about what ends are just and unjust.

(F3) Judgments about what is just are seen to be supported by good reasons.

(F4) The plan of life, considered as a whole, has a dominant theme—we have "an identifiable character that shows itself in [our] deliberations and that persists over time."[249]

When plans are dictated by rational choice (F1), they contain the ends they do because those ends are prescribed by principles taken to be requirements of reason. Clearly the extent to which plans are dictated by rational choice is a matter of degree. If the only requirement of reason we acknowledge is instrumental rationality, for example, then a very wide amount of scope will be left for purely preferential choice. At the other extreme, we could acknowledge principles that held that only one possible plan of life would be in accordance with reason. Many sets of principles will sit somewhere in between, ruling out some plans but leaving some scope for preferential choice.

Significantly more scope will be left for purely preferential choice when principles of justice are not acknowledged as supremely regulative. This means that (F3) will be lacking, as Weithman explains:

> Suppose I must do without some resources because justice demands the satisfaction of others' ends, and suppose I know those ends to have been adopted on the basis of others' purely preferential choice. Then the claims that are to be honored in preference to mine will seem arbitrary. So too will the verdict of justice that supports honoring those claims.[250]

(F2) may also be lacking, because the distribution of preferences may change over time such that what is permissible at one point may later be prohibited. In order to maintain consistency between the ends included in our plan and what is judged to be right, we would have to keep changing our plans to fit with the overall distribution of preferences in our society. Therefore, we would not be able to plan over the long term to have ends consistent with justice, as the content of such a consistent plan would always be liable to change. We would not, therefore, be able to achieve (F2)

---

[249] I take these features from Paul Weithman, *Why Political Liberalism?* 211-212. Weithman extracts them from Rawls's comments on the unity of the self (*A Theory of Justice*, §85), among other places—they do not appear in this list form in the text of *Theory*.

[250] *Why Political Liberalism?* 213.

without sacrificing (F4), as if we constantly modify our plan to track the distribution of preferences in our society, we will not be able to live out a plan that shows that we have "an identifiable character that shows itself in [our] deliberations and that persists over time".[251]

We can eliminate these problems is by framing our plan of life in accordance with principles of justice acknowledged as supremely regulative. When we acknowledge such principles, the scope for purely preferential choice is more limited, so we do better with regard to (F1). More importantly, the remaining scope that purely preferential choice does have does not influence what is just in the way that was problematic above. Further, changing preferences over time will also not influence what is just, so we can achieve (F2) without sacrificing (F4). We might say that when plans are framed in this way the content of justice is insulated from individual preferences, such that what is just is not influenced by what citizens happen to prefer. When justice is insulated in this way, it makes it possible for us to plan our lives such that they exhibit unity.[252]

With this account of rational unity, we now have before us a full picture of the desire to express our rational nature that the KCA appeals to. So according to Rawls, the desire to act on the principles of justice as fairness, the desire to be a just person, the desire to act on the principles that would be chosen in the OP, and the desire to express our rational nature all amount to the same thing.[253] The desire to express our rational nature just is the desire to act on the principles that would be chosen in the OP, as the OP is designed to express our rational nature. And since, as I am assuming for now, the two principles of justice as fairness would be chosen in the OP, the desire to express our rational nature is also equivalent to the desire act on these principles. And

[251] Ibid., 212.

[252] An odd feature of this account of the interest in rational unity, as specified by Weithman, is that what seems to matter is not at all whether we acknowledge principles of justice as supremely regulative, but what the content of those principles is. There is no reason why we could not acknowledge the principle of utility as supremely regulative of our rational plans. However, acknowledging it as such would not allow us to satisfy this interest in the unity of our plan if we have such an interest. The problem with the principle of utility from the point of view of this particular interest is that leaves the demands of justice liable to change over time as the contents of citizens' desires and plans change. But this, it seems, is clearly a problem generated by the content of the principle of utility rather than by whether or not it is acknowledged as supremely regulative. However, since nothing hangs on this issue for the argument of this chapter I will set it aside.

[253] Ibid., 503.

finally, since these are principles of *justice*, an alternative way to describe all of these desires is as a desire to be a just citizen.

B.

Showing that these desires all have the same content is an important step in the KCA. Recall that the aim of the KCA is to show that justice as fairness satisfies the stability requirement, and thus the publicity requirement, by arguing that that citizens who grew up in a society well-ordered by the principles of justice as fairness would have a desire that is practically identical with the desire to be a just person, and that this is a desire they can only satisfy by accepting the principles of justice and treating them as regulative of their other desires. However, though an important part of establishing this conclusion is identifying a desire that has the same content as the desire to act on the principles of justice as fairness, the other more important part is to show that the citizens of the WOS would actually acquire this desire and choose to treat it as decisive.

Let me now take a look at what Rawls has to say about this next part, beginning with the acquisition of the desire. He offers a general account of how the desire to be a just person might arise in those who grow up in the WOS, via a three stage process of moral development. Those stages, in outline, are as follows:

> THE MORALITY OF AUTHORITY: Given that family institutions are just, and that parents love the child and manifestly express their love by caring for his or her good, then the child, recognising their evident love of him or her, comes to love them.
>
> THE MORALITY OF ASSOCIATION: Given that a person's capacity for fellow feeling has been realised by acquiring attachments in accordance with the first stage, and given that a social arrangement is just and publicly known by all to be just, then this person develops ties of

> friendly feeling and trust toward others in the association as they with evident intention comply with their duties and obligations, and live up to the ideals of their station.
>
> THE MORALITY OF PRINCIPLES: Given that a person's capacity for fellow feeling has been realised by their forming attachments in accordance with the first two stages, and given that a society's institutions are just and are publicly known by all to be just, then this person acquires the corresponding sense of justice as they recognise that they and those for whom they care are the beneficiaries of these arrangements.[254]

Although these three stages gloss over many details that would undoubtedly be included in the full story, they provide a sketch of how a citizen of the WOS could gain the desire to be a just citizen. At each stage their desires are transformed by the effects of the living in a society whose institutions and members conform to certain principles, until they themselves develop a desire to act on those principles. And though this sketch refers to the acquisition of the desire to be just, if we grant the first step of the KCA then this desire has the same content as the desire to express our rational nature, so it can likewise be used show how a citizen of the WOS could gain the desire to express their rational nature.

A natural first response to this sketch might be to say it is too speculative to be the basis for an argument for a conception of justice. However, though it is indeed a speculative sketch I do not think that this should be a cause for concern. It seems plausible to suppose that different processes of moral development can bring about desires to follow principles or realise ideals of many different kinds. It is evidence for this claim that the world contains persons who are committed to, and act on, a wide variety of moral principles and ideals. If there were some natural law such that a process of moral development could only lead to desires to act on principles within a narrowly specified range, then this evident pluralism could not arise. Given that it does arise, I

[254] *A Theory of Justice*, 429-430.

think we should suppose that it is possible for processes of moral development to bring about desires to follow a wide range of principles and ideals. Therefore, I do not think we should doubt this sketch solely for its speculative nature. Perhaps there is some special reason to doubt that the desire to express our nature as rational beings could not come about via a process of moral development, but I think the burden of proof is on those who doubt that it could to explain why.

This defence of the account might give rise to another worry, however. If desires to follow all sorts of principles and ideals can be brought about via processes of moral development, then the fact that a desire to act on Rawls's principles of justice—or any other principles for that matter—can be brought about in this way will not count very strongly in favour of those principles. After all, if I am right then there is little that is special about this fact: it is simply the fact that Rawls's conception of justice is one of a wide range of conceptions of justice for which this desire can be generated. And if there are a wide range of conceptions of justice for which this desire can be generated, it might seem that there are a wide range of conceptions of justice that can satisfy EC.

I think it is right that the fact that a desire to act in accordance with particular principles of justice could be acquired via a process of moral development does not count strongly in favour of those principles, though I do not think that this is at all troubling. Recall that EC takes the moral principles that satisfy the publicity requirement to be correct because those are the principles we would accept if our moral competence was operating free of distortion. Satisfying the publicity requirement involves showing both that a desire to act on the principles would come about in the WOS, and that the citizens of the WOS would freely accept these principles, thus choosing to maintain their desire to act on them. Though showing that the desire to act on the principles could be acquired is a necessary component of satisfying the publicity requirement, the normative force of that requirement stems principally from its second component: that citizens of the WOS freely choose to maintain their desire to act on them. Therefore, the fact that desires to follow all sorts

of principles and ideals can be brought about via processes of moral development does not give us reason to think that they would thereby satisfy EC.

C.

Let me now move on to what the KCA has to say about how the citizens of the WOS would weigh the desire to express their rational nature, assuming that they have acquired it. Establishing that the desire to express our rational nature would be considered to be sufficiently weighty in the WOS is a fundamental part of the KCA. If the KCA were to only establish that this desire would have some positive weight but would sometimes or often lose out to other desires when citizens of the WOS are deciding how to act, then it will not have established that justice as fairness satisfies either the publicity requirement or the stability requirement. The satisfaction of either of those requirements needs an effective desire to act on the principles of justice, not just some desire.

Rawls attempts to establish the weight of the desires held by citizens of the WOS by considering how those desires tend to move us in world as it is.[255] If, in the world as it is, we find that the desire to express our rational nature is decisive, then given that the conditions of a WOS are more favourable than the conditions of the world as it is, we can conclude that it would also be decisive for citizens of the WOS. Do persons in the world as it is take the desire to express their rational nature to be decisive? It might seem to be fairly obvious that they do not. After all, the desire to express our rational nature is a desire to act on the principles that would be chosen in the OP, and, I am assuming, the principles of justice as fairness would be chosen in the OP, and it is very clear that persons in the world as it is do not have a decisive desire to act on these principles. If they did we would be inhabitants of a world in which there was full compliance with Rawls's principles, yet we are not.

Paul Weithman considers this difficulty and gives the following reply on behalf of Rawls. The OP is choice scenario in which principles of justice are chosen *for a WOS*: a society which

[255] *A Theory of Justice*, 501; Weithman, *Why Political Liberalism?* 161.

exhibits full compliance with principles of justice. That we do not have a decisive desire to live according to principles chosen *for such a society* does not mean that we do not have a desire to express our nature in the world as it is. The desire to express our nature is not a desire to regulate our lives according to the principles that would be chosen for an ideal society, regardless of what people in our non-ideal society are actually doing. In the world as it is, the object of the desire to express our rational nature would be to live according to the principles that would be chosen if our nature were the decisive determining element, but not supposing that we are choosing for a society that exhibits full compliance but rather exhibits whatever degree of non-compliance we face.[256] Therefore, the desire to express our nature amounts to something quite different in the world as it is to what it would amount to in the WOS. Our world is unlike the WOS in numerous ways: we have to respond to the fact that others behave unjustly; that we do not have sufficient information about how others are behaving and that we do not always have information that would help us make a choice of what principles to govern our lives by. In spite of this, Weithman states, we still want to live according to the principles that would be chosen if our rational nature was the decisive determining element in our decision, but these are principles modified for the non-ideal conditions in which we find ourselves.[257] This point does not threaten the claim that in the circumstances of a WOS the desire to express our rational nature has the same content as the desire to live according to the principles that would be chosen in the OP. The reason we do not have a desire to live according to principles for a society of full compliance here and now is because to do so would leave us liable to losses from the non-compliance of others. In the WOS, where, by stipulation, everyone is complying and we are aware of their compliance, this problem would be gone, and so regulating our plans in accordance the principles that would be chosen in the OP would then be wholly appropriate.

[256] Ibid., 198.
[257] Ibid.

D.

That concludes my description of the KCA. Does it succeed? If not, what is wrong with it? Rawls thought that it failed, and his commentators have tended to agree with him. He held that the argument of *A Theory of Justice* relied on a comprehensive philosophical doctrine that was incompatible with the kind of pluralism that would be found in a WOS.[258] And he stated that all of the subsequent recasting of his view in *Political Liberalism* was motivated by problems with his argument for the stability of justice as fairness in *Theory*. The KCA is fundamental to that argument, and it appears to rely on a comprehensive philosophical doctrine. Therefore, it appears his official view is that the KCA fails because of its reliance on a comprehensive philosophical doctrine.

However, there is disagreement among commentators about what precisely was wrong—and what Rawls thought was wrong—with the argument for the stability of justice as fairness in *Theory*.[259] Since I want to know what, if anything, satisfies the publicity requirement, and if the KCA succeeded it would show that justice as fairness satisfies that requirement, what exactly is wrong with the KCA is an important question for my purposes. If it is a problem that can be fixed, we will need to know how, and what influence, if any, this has the on the content of the conception of justice that it shows to satisfy the publicity requirement. If it is a deep and intractable problem, then we may have return to my conception of the circumstances in which our moral competence operates and see if they need to be revised.

As before when I was specifying EC, in considering what if anything is wrong with the KCA I will not focus primarily on trying to defend a reading of Rawls's text. I will focus instead on analysing the KCA on its merits, leaving open the possibility that the reason the argument in fact fails is different from Rawls's own reason for thinking it fails. As I have already suggested, I doubt that the KCA is flawed with respect to Rawls's account of how the desire to express our rational nature is acquired as part of a process of moral development. As I argued above, I suspect

---

[258] *Political Liberalism*, xvi.

[259] Compare the accounts in: Brian Barry, "John Rawls and the Search for Stability," *Ethics*, 105(4), 1995, 874-915; Samuel Freeman, "Congruence and the Good of Justice"; Gerald Gaus, "The Turn to a Political Liberalism," in *A Companion to* Rawls; and Weithman's *Why Political Liberalism?*

a wide range of desires can be acquired in this way and I see no special reason to think that this desire in particular could not be. Therefore, I think that any problem with the KCA will stem from the particular desire that has been identified or with the claim that this desire would move citizens of the WOS decisively. These two issues are intimately related. The prospects for the KCA or an argument like it depend on the possibility of identifying a desire that is decisive in this way, and the plausibility of any claim that a particular desire would move citizens of the WOS decisively depends entirely on the content of that desire. So there is essentially one task here: identifying a desire that citizens of the WOS would treat as decisive.

6.4. Kantian Autonomy and the Desire to Express our Rational Nature

My ultimate aim, as I stated in the introduction, is to show that an argument with the form of the KCA can be used to show that a particular conception of justice satisfies the publicity requirement. I will thus begin my evaluation of the KCA by examining what if anything is attractive and plausible about it, and then go on to consider its limitations. Why might one think that the desire to express their rational nature is one that citizens of the WOS would take to be decisive?

To consider this, let me now take a look at the desire in a little more detail. Once again, to see ourselves as having a rational nature is to see ourselves not as identical with any of our particular ends or interests, but rather as someone who can revise and alter their ends. The desire to express this nature is a desire to act on the principles that would be chosen if it were the decisive determining element. One attractive feature of this desire is that it expresses the idea that our particular ends are always *revisable*. That we must accept that our chosen ends may turn out to be mistaken, and thus require either modification or potentially abandonment, is an incredibly difficult claim to deny. The reason for this is that to not see our ends as revisable in this way is to commit ourselves to holding that our judgment in choosing our ends was infallible.[260] We must either hold

[260] Allen Buchanan adeptly draws attention to this point in his "Revisability and Rational Choice," *Canadian Journal of Philosophy*, (5)3, 1975, 395-408. However, his aim is distinct from mine here: he appeals to the ideal that a rational person must seem themselves as fallible chooser of their ends in order to argue for the conclusion that a preference for primary goods is uniquely rational.

that the ends we have chosen for ourselves are in principle revisable, or that they are not. But to hold that they are not we must think that our judgment of their relative merits is infallible, and this, it seems, would clearly be to make a mistake.

I think this is a strong attraction of the desire to express our rational nature and a reason one might think it could be a desire the citizens of the WOS would all take to move them decisively. Instead of appealing to some particular end which we might for good reason doubt could be shared in a liberal society like the WOS, it appeals to an identity that purports to be prior to all of our other ends: our identity as a fallible chooser.

However, for citizens of the WOS to have a decisive desire to express their rational nature it is not enough that they all agree that they are fallible choosers of their ends. The desire amounts to much more than this, for it is a desire to act on the principles that would be chosen if our rational nature were the decisive determining element of our choice. Our rational nature plays this determinative role when it dominates, as it were, our particular ends themselves, as in the choice of principles in the OP. The design of the OP is such that our rational nature must be determinative, as being behind the veil of ignorance the parties have no knowledge of their ends, and so are unable to weigh their desire to satisfy them against the desire to express their rational nature. The citizens of the WOS, however, are in a markedly different position: they have ends they want to pursue and they know what those ends are. In choosing what principles to act on, they are not forced like the parties to the OP to set aside the particular ends they want to pursue. Even acknowledging that they are fallible choosers of their ends, it not clear that they must thereby set aside those ends in choosing what principles to act on. In acknowledging that you might be mistaken, the rational response is usually to still to act on principles that maximise the fulfilment of your ends, but with the weight accorded to each end adjusted in accordance with your judgment of the probability that it will later be revised or rejected. Since citizens of the WOS will have, we can assume, a wide variety of ends to which it will be rational for them to accord a range of weights,

it not clear that reasoning this way will lead them to act on the principles that would be chosen in the OP.

Let me use an example to make this more concrete. Consider a citizen of a society that is well-ordered by the two principles of justice as fairness. Suppose that this citizen has been brought up in this society and acquired, via the process of moral development sketched above, a sense of justice informed by these two principles. Suppose further that she also has the related desire to express her rational nature. She sees herself ultimately as a being who can revise and alter her final ends, not as identical with any of her ends in particular, and wants to express her nature as a being of this kind. Now, instead of asking the general question of whether it would be rational for her to maintain this desire, I want to consider a scenario in which her pursuit of her interests and her sense of justice appear to give different recommendations:

> *Constitutional Amendment*: The governing political party proposes an amendment to the constitution that would revoke the basic liberties of a small minority of the population. This amendment would further the interests of the current majority, of which Betty is a member and whose interests she shares. The probability that the amendment will have the desired effect with respect to the long term interests of the majority is very high, and Betty has no reason to suppose that that she might somehow become a member of the minority or lose out from the amendment in some other way. Given all this, we can say that there is a clear sense in which this amendment would seem to be of great private benefit to Betty.

I will refer back to this case at a number of points in what follows as it illustrates the conflict between what is a plausible requirement of justice and the point of view adopted by citizens of the WOS when they bracket their desire to be just for its own sake and reason on the basis of their other interests. If it could be shown that Betty would choose not to vote in favour this amendment

when reasoning in this way, we would have a good starting point for showing that an acceptable conception of justice satisfies the publicity requirement.

So, what would Betty make of this amendment? Given that she has a sense of justice regulated by the two principles of justice, and those principles hold that to revoke the liberties of the minority would be severe injustice, she would strongly oppose it. However, when she abstracts from her sense of justice and considers the question from the point of view of whether it would further her other interests it seems natural to suppose that—given the stipulations of the example—the amendment is in her interests and thus something she ought to support. What about her desire to express her rational nature, which we are assuming she has also acquired? The attractive feature of that desire that I have drawn attention to cannot be used to show that this desire would be decisive in this instance, for the reason I have already given. Supposing Betty, because she is reasoning well in circumstances of full information, accepts that she is a fallible chooser of her ends, it does not follow that she must adopt the constraints of the OP in deciding what principles to act on. She can do justice to her fallible nature simply by acting on principles that adjust the weights of her particular ends accordingly.

In setting out the KCA Rawls argues that performing these adjustments is not sufficient to express our rational nature. In concluding the argument, he writes that to not act on the principles that would be acknowledged in the OP is "to not achieve for the self free reign but to give way to the contingencies and accidents of the world".[261] In order to properly understand this claim, and see the full explanation of why Rawls held that in order to express our rational nature we must act on the principles that would be chosen in the OP, we need take a detour via the Kantian interpretation of justice as fairness.

Recall that the desire to express our rational nature is the desire to act on the principles that would be chosen if our rational nature were the decisive determining element of our choice.

[261] *A Theory of Justice*, 503.

This desire draws on the claim that the OP can be given a Kantian interpretation. On Rawls's reading, Kant held that:

> A person is acting autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being. The principles he acts upon are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things that he happens to want. To act on such principles is to act heteronomously.[262]

For Betty to adjust the weight of her ends in the manner I have suggested, and then to conclude that she ought to support the constitutional amendment would thus be for her to fail to act autonomously. Citizens of the WOS fail to act autonomously, on this view, whenever they accord their particular ends determinative weight in their decision of what principles to adopt: to act autonomously is act on the principles that would be chosen if such contingent things as our particular ends and desires were not allowed to influence our choice.

The desire to express our rational nature is thus a desire for a particular kind of autonomy. This is the autonomy that we achieve when the principles on which we act are entirely uninfluenced by the contingent features of the world. A desire for this kind of autonomy is one we can only satisfy by acting on the principles that would be chosen in the OP, as the OP abstracts away from all of these contingencies. Now, if it were the case that Betty had a decisive desire for this kind of autonomy, then she would indeed reject the constitutional amendment. And if every citizen of the WOS was moved decisively by this desire, then the KCA would succeed: everyone would want to comply with the principles of justice as fairness, even when they abstracted from their desire to be just for its own sake.

What grounds do we have for supposing that the citizens of the WOS would be moved decisively by this desire? Weithman's suggestion, which I set out above, was that in the world as it is people want their rational plans of life to be governed by principles that would be chosen if their rational nature was the decisive determining element in their decision, and since the circumstances of the WOS are more favourable than the circumstances of the world as it is, we should thus expect

[262] Ibid., 222.

that citizens of the WOS will also want their plans of life to be governed in this way. Now that we have a full picture of what the desire to express our rational nature needs to amount to for the KCA to succeed, we can properly consider this suggestion. We can say, I think, that rational persons in the world as it is would not see themselves as identical with their particular ends or interests, for the fallibility-based reason I gave above. However, we cannot say, without further argument, that persons in world as it is tend to have a decisive desire to avoid Kantian heteronomy. Though some may want to abstract from their particular desires in this way, others will express their rational nature simply by acting on principles that adjust the weights accorded to their particular ends according to their assessment of the probability that these are sound ends to pursue. If that is right, then there are no firm grounds for thinking that the citizens of the WOS would be moved decisively by the desire to express their rational nature.

It is here, then, that the KCA can be said to rely on a comprehensive philosophical doctrine that there are reasons to doubt would be the object of convergence in the WOS. The desire to be autonomous in this way is comprehensive: to have it is to have a desire that regulates the entirety of one's plan of life. It is also a desire that is in conflict with a range of comprehensive philosophical doctrines and religious views that do not hold that this kind of heteronomy is to be avoided. Of course, it is not troubling in and of itself that the desire to express our rational nature belongs to a particular comprehensive philosophical doctrine and would be rejected by those who hold others. If it could nonetheless be shown that the desire would move citizens of the WOS decisively, then its comprehensiveness would be irrelevant. However, the KCA does not appear to provide grounds for thinking that this would be the case, as Rawls later accepted.

6.5. The Argument from the Mutability of Social Positions

The conclusion of this discussion of the KCA is that though the argument has the attractive feature of appealing to a desire expressive of our fallibility, it fails to explain why citizens of the WOS would have a decisive desire to avoid heteronomy. I will now develop an argument that is

structurally analogous to the KCA, but appeals to a different desire. I will refer to this desire as the desire to express our fallibility, because I intend to continue to make use of the thought that citizens of the WOS accept that they are fallible choosers of their ends. My argument has two aims. The first is to outline this desire and show that satisfying it requires them to act on particular principles of justice, such that if our citizen in *Constitutional Amendment* were moved by decisively by this desire, they would reject the amendment. The second is to offer grounds to suppose that this desire would move citizens of the WOS decisively regardless of their particular ends or interests.

The starting point of the argument is thus that citizens of the WOS would see themselves as fallible choosers of their particular ends. The case for this claim is simple. In the world as it is, when we are thinking clearly, we would not deny that we are fallible in this way: to do so would be absurd. In the ideal conditions of the WOS there is no reason to suppose things would be any different. However, as I pointed out in my discussion of the KCA, seeing ourselves as fallible in this way does not seem to give us reason to take up the constraints of the OP in deciding how to act, for it is only to accept the fairly weak claim that in principle the ends that we have chosen for ourselves may turn out to be mistaken.

I now wish to consider this claim about fallibility alongside the special context of the institutions that make up the modern state, or what Rawls called the basic structure of society. These institutions include the political constitution, the system of property rights, the structure of the economy and the family, as well as any other main social and political institutions.[263] The desire to express our fallibility applies specifically to this context. Like the desire to express our rational nature, it is a desire to act on the principles that would be chosen if our nature as fallible choosers of our ends were the decisive determining element. However, the scope of the desire is such that it is only a desire to act on such principles for decisions relating to the institutions that make up the basic structure of society: having it implies nothing about the principles citizens want to act on in other domains. Since this desire is a circumscribed version of the desire to express our rational

[263] *Political Liberalism*, 11-12.

nature, it could also be described as a desire for a particular kind of autonomy. Those who have this desire do not want their choice about what principles to act on in political life to be determined by the particular ends or desires that they happen to have. They want instead to act on the principles that they would accept if they abstracted away from these contingencies.

I will now present an argument that aims to show that the desire to express their fallibility would move citizens of the WOS decisively if a conception of justice with certain features is realised. For the first step of this argument I need to dispel a quite natural way about thinking about social positions in the WOS. This is that we might think of citizens of the WOS as emerging from behind the veil of ignorance into a particular social position, the prospects of which will be defined by the principles that were chosen in the OP, and that this citizen of the WOS will continue to occupy this social position for the entirety of their life. It is natural to think of the social position that citizens occupy in the WOS, whether as the least advantaged or otherwise, as to some degree fixed. However, how well our social position fares will depend to a significant degree on the particular ends with which we find ourselves. Take a familiar example: if a citizen of the WOS emerges from the behind the veil of ignorance with a religious conception of the good that they are not permitted to practice (as no principle guaranteeing freedom of religion was chosen by the parties), then that citizen may face intolerable costs.[264] Given that how well a citizen of the WOS's social position fares can depend on her particular ends in this way, it is of great significance that she recognises that she is a fallible chooser of those ends. For, given this, she does not see herself as identical with those ends and takes them to be in principle revisable. That she might revise her ends opens up the possibility that her social position might change dramatically if she makes a particular revision. For example, she may revise her ends such that she becomes a part of a religious minority who are not afforded basic liberties. This possibility generates a reason to uphold a set of basic liberties for all (and perhaps much more) that citizens of the WOS would still have when they abstracted away from their desire to be just for its own sake. Acting on principles that would make

[264] Rawls, *A Theory of Justice*, 180-181.

the prospects of another social position intolerable is something that has a cost from the self-interested point of view: it is possible that they could come occupy that position, and thus experience the burdens associated with it.

Returning to *Constitutional Amendment*, the foregoing reasoning provides Betty with some reason to refrain from endorsing the proposal. Taking away the basic liberties of the minority would be to negatively impact a social position that she herself could end up occupying. But for all I have said so far, this reason might be accorded very little weight in her reasoning about what principles to act on. Betty may grant that there is some positive probability that she will end up in this position, but consider it to be sufficiently small that making the minority position intolerable is a risk worth taking. Provided that what she has to gain is of sufficient benefit to her to outweigh the risk that she ends up as an occupant of that social position, maximising the fulfilment of her interests will lead her to support the amendment.

Now, one way in which this reasoning could support Betty's decision to reject the constitutional amendment is if her reasoning used or approximated what is known as the maximin decision rule. If her reasoning used or approximated that principle, then she would make her decision about what principles to act on by considering the worst potential outcome under each principle, and acting on the principle for which the worst outcome is as good as possible. In this case, she would not conclude that making the position of the minority intolerable is a risk worth taking.

This depends on Betty's reasoning using or approximating the maximin decision rule. However, there are strong reasons to doubt that it would be rational for her to reason in this way. As a general principle for making decisions, the maximin principle is generally thought to be implausible. Consider the following famous critique from John Harsanyi:

> Suppose you live in New York City and are offered two jobs at the same time. One is a tedious and badly paid job in New York City itself, while the other is a very interesting and well paid job in Chicago. But the catch is that, if you wanted the Chicago job, you would have to take a plane from New York to Chicago (e.g., because this job would have to be taken up the very next day).

Therefore there would be a very small but positive probability that you might be killed in a plane accident.[265]

In this case the maximin principle tells you to take the job in New York City due to risk of the plane accident, but this seems absurd: the small but positive probability of an accident occurring does not render choosing the job in Chicago irrational.

As Rawls has argued, however, though the maximin principle is clearly unfit as a general principle of rational choice, there are grounds for making our reasoning approximate it in certain conditions. Namely, if three conditions hold to a sufficient degree: (1) if we have no reliable basis for estimating the probability of ending in the worst off social position; (2) if the worst outcomes are intolerable and must be avoided; (3) if we care little for what might be gained above what can be guaranteed, because what can be guaranteed is satisfactory.[266] When these three conditions hold, then applying the principle that we ought to maximise the fulfilment of our interests will lead us to approximate the maximin rule. If we care little for what we could gained by taking the risk of making the worst possible outcome intolerable, and we cannot reliably assign a probability to that risk, it will be irrational to take it. As should be clear, these conditions do not hold in Harsanyi's example. We can with some confidence assign a probability to the possibility of our plane crashing, and the difference in desirability between the two jobs as Harsanyi describes them is significant.

That there are conditions in which it would be rational to reason in a manner that approximated the maximin rule gets us to the following conditional:

> If a conception of justice makes the circumstances of the WOS such that, for the decision of what principles to act on with respect to the basic structure of society, approximating the maximin rule is rational, then that conception of justice satisfies the publicity requirement.

[265] "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review*, 69(2), 1975, 595.
[266] For these conditions, see *Justice as Fairness: A Restatement*, 97-98.

This is because I have argued that Betty has some reason to maintain the basic liberties of her co-citizens, and in these special conditions that reason would move her decisively, as her reasoning would approximate the maximin rule. Therefore, if there is a conception of justice that would bring about the conditions that make approximating maximin rational in the WOS, it would satisfy the publicity requirement.

We can state this argument more fully as follows:

(1) Citizens of the WOS accept that they are fallible choosers of their particular ends, and thus that they may come to modify or abandon them.

(2) How well particular social positions in the WOS fare depends on the ends of their occupants.

(3) It is possible that a citizen will modify her ends such that she comes to occupy a less advantaged social position.

(4) The circumstances of the WOS are such that the following conditions hold to a sufficient degree to make approximating the maximin rule rational.

(a) citizens have no reliable basis for estimating the probability that a change in their ends will lead them to occupy a less advantaged social position.

(b) the worst possible outcome, if the status quo is modified and a citizen's ends change, would be intolerable.

(c) citizens find the status quo satisfactory, and care little for what they might gain by changing it to their personal advantage.

(5) Therefore, a citizen of the WOS would not desire to act on principles that aim to change the status quo to her personal advantage.

I will call this the argument from the mutability of social positions. One important advantage that this argument has over the KCA is that at no point does it assume or rely on the claim that citizens of the WOS share anything more than a commitment to their fallibility as choosers of their ends. If this argument is sound, and there is a conception of justice that generates the conditions

presented in premise (4), then it can be used to show that that conception of justice satisfies the publicity requirement.

To see this, recall that in order to show that a conception of justice satisfies the publicity requirement we need to show that citizens of the WOS would acquire a desire to act on its principles as a result of their upbringing, and that they would take this desire to be decisive even if they abstracted away from their desire to be just for its own sake and reasoned from a self-interested point of view. I have already argued that the first step is easy enough to satisfy. Desires to act on a number of different principles and ideals can be brought about via processes of moral development. I therefore think it is plausible to suppose that citizens of the WOS could come to have what I have called a desire to express their fallibility: a desire to abstract away from their particular ends and characteristics when they are making decisions in a political capacity, as an acknowledgement of their fallibility as choosers of their ends. The argument from the mutability of social positions could allow us to secure the second step. It supplies a reason for citizens to abstract away from their particular ends and characteristics in these circumstances that they would take to be decisive, if conditions described in (4) are in place.

I think there are two major difficulties with this argument that will need to be addressed if it is to succeed. The first relates to the scope of the claim that how well particular social positions fare depends on the ends of their occupants. The second is whether any conception of justice could be such that it realises the conditions described in (4) to a sufficient degree.

The first difficulty can be put in the form of an objection, which runs as follows. The claim that a citizen's conception of her ends might change such that she comes to occupy a less advantaged social position is limited in various ways. It holds for the cases where social positions are specified in terms of particular conceptions of the good where it is possible to come to occupy that position by coming to affirm that conception of the good. It appears not to hold, however, for cases in which social positions are specified in terms of natural characteristics or social meanings such that one cannot come to occupy them simply by changing one's view or practices. To take an

example, suppose that we modify *Constitutional Amendment* such that instead of Betty considering whether the basic liberties of a religious minority should be revoked, she is considering the status of citizens over six feet tall. If Betty is five foot five, the position of 'citizen over six feet tall' is not one that she could come to occupy in her lifetime. Therefore, she would not have a reason to refrain from acting in a manner that renders this social position worse off, or even intolerable, that stems from the possibility that she herself could come to occupy it. Or suppose instead that we modify the case such that Betty is considering voting to revoke the rights and liberties of a group that one cannot be a member of unless one is the descendent of a member of that group. Assuming Betty is not such a descendent, the social position of a member of such a group will not be one that she could come to occupy in her lifetime. So, again, she would not have a reason to refrain from acting in a manner that renders this social position worse off, or even intolerable, that stems from the possibility that she herself could come to occupy it. This reasoning will likewise hold for any case in which social positions are specified in terms of characteristics one cannot take on simply by changing one's view or practices. This objection therefore threatens to leave premise (2) of the argument too weak to deliver the conclusion that Betty would choose to maintain a status quo that included basic liberties for all citizens, regardless of their natural characteristics.

We should not, however, think of how well a particular social position in the WOS fares as depending solely on what is in what might be called the narrow self-interest of the individual who occupies it. When citizens of the WOS abstract away from their desire to be just for its own sake, this does not lead them to only care about themselves. Even with the desire to be just for its own sake out of the picture, citizens may still care about other people such as friends, family and other loved ones. It is an open question how far reaching these other-regarding desires will be for a citizen of the WOS, but it is possible that they stretch rather widely. Acknowledging that there will be these other-regarding desires opens up the possibility that citizens of the WOS will have reasons to refrain from making other social positions worse off, even if they themselves will never come to occupy them. So, to return to the case of Betty considering whether citizens over six feet

tall ought to be afforded basic rights and liberties, it does not follow from the fact that she will never occupy this position that she has no reason to protect those rights and liberties. If she currently cares for someone with that natural characteristic, or foresees that she may come to care about someone with it in the future, then this will provide her with some reason to protect the rights and liberties associated with its social position. An acknowledgment of her fallibility is relevant here too, because how widely a citizen's other-regarding desires will stretch depends on their particular conception of the good. Thus in acknowledging that they are fallible choosers, citizens must accept not only that their conception of the good might change such that they themselves come to occupy other social positions, but also that it might change such that they come to care about the occupants of social positions that they do not currently care about. Both of these are ways in which their own position might change for the better or for the worse by their own lights.

The second and more fundamental difficulty is this: is there a conception of justice that would bring about the conditions that premise (4) refers to in the WOS? I will now consider each of these conditions with that question in mind.

(a) *That there would not a be a reliable basis for estimating the relevant probabilities*

It may seem somewhat natural to think that we have a firm basis for estimating the probability that our conception of the good will change considerably, and thus that citizens of the WOS would have an even firmer such basis given that they are in conditions more ideal than ours. This line of reasoning might begin by claiming that, once in adulthood, people rarely change their conception of the good radically, such as by converting into or out of a religious view. And regardless of how often this occurs, we have firm grounds for estimating the probability that it will happen to us: all else equal, it is as likely to happen to us as it is to others. So if we are adherents of conception X, and on average N% of that conceptions adherents convert to conception Y in their lifetimes, we can proceed on the basis that there is a similar chance we will likewise convert in this way, and

repeat this reasoning for any other possibilities. We can thus assign rough probabilities to each option, and thereby to the possibility that a change in our conception of the good will cause our social position to change for the worse.

As we will see as I go on to discuss the other two conditions, the extent to which any one of the conditions needs to hold in order for it to be appropriate for reasoning to approximate the maximin rule depends on the extent to which the other two also hold. With respect to this condition, it is clear that it at least needs to hold enough for citizens to be unable to securely place a very low probability on their conception of the good changing dramatically. If they were able to do this, then their situation would be like that of Harsanyi's job seeker deciding whether to fly to Chicago.

One possible route that could ensure that this condition is satisfied would be to systematically misinform the citizens of the WOS about the relevant probabilities. However, this would leave the argument from the mutability of social positions incapable of being combined with the publicity requirement, as that requirement requires citizens of the WOS to be in conditions of full information. Since my interest in the argument from the mutability of social positions is ultimately an interest in working out what conception of justice would satisfy the publicity requirement, any route to satisfying condition (a) will need to respect the full information requirement.

I will now present an alternative route based on two claims, one empirical and one appealing to the special conditions of the WOS. The first claim is that the line of reasoning I set out above errs in claiming that in the world as it is people rarely change their conception of the good radically, such as by converting into or out of a religious view. In fact, recent evidence from the US suggests that religious conversions are relatively commonplace with approximately half of all adults changing their religious affiliation at least once during their lives.[267] Nor does the evidence suggest a one-way street of changes away from affiliation with one of the major religious identities

[267] See the Pew Research Group's "Faith in Flux: Changes in Religious Affiliation in the U.S." *The Pew Forum on Religion and Public Life*, 2009, 1-4.

toward becoming unaffiliated: approximately half of those raised unaffiliated later become religiously affiliated.[268] Further, many change their affiliation more than once over the course of their lifetime.[269]

We might be minded not to given much weight to this evidence. After all, citizens of the WOS are in more ideal conditions for making judgments about what conception of the good to follow, and we might think that this will bring about significantly greater stability in their affiliations. However, I think there are strong grounds to hold that citizens of the WOS would see the plurality of conceptions of the good adhered to in their society in a different way to that in which many citizens of the world as it is see pluralism. As we have seen, in the WOS citizens are symmetrically situated: they are all in circumstances where their competences are operating free of distortion. Knowing that both they and their co-citizens are rational and in ideal conditions, they would not see the views held by their co-citizens as alien doctrines. They would see them instead as doctrines that they themselves could have ended up holding. Given the circumstances of the WOS, disagreements between its citizens are explained by the burdens of judgment. Knowing this fact about the explanation for their continued disagreement members of the WOS will also know that if these burdens had operated at all differently, they could have come to hold a different conception of the good. In this sense, then, they would not see the conceptions of the good held by their co-citizens as clearly or deeply mistaken, in the way that persons in the world as it is may well see some of the alternative doctrines that they encounter. Instead, they would see them as close possibilities. Note that this reasoning does not depend on the citizens of the WOS being misinformed in any way. If they are, as we are supposing, in circumstances where their competences are operating free of distortion, then this claim about the status of alternative conceptions of the good is quite true.

I believe this reasoning rules out citizens of the WOS assigning a very small probability to the possibility of their conception of the good changing radically. Many people in the world as it is change their conception of the good. And in circumstances where the alternative conceptions held

[268] Ibid., 17.
[269] Ibid., 1.

by their co-citizens appear to citizens of the WOS as close possibilities, there are no grounds to suppose such changes would be less likely. Indeed, they may be even more common. I also think that we can say that citizens of the WOS would not have a reliable basis for assigning a probability to the possibility of their conception of the good changing radically. The set of possible conceptions of the good is to some degree in flux, such that they cannot say for certain that there are a particular number of possibilities. And even though they can observe the extent to which others who are similar to them end up changing their view, they cannot know to what extent they themselves might constitute a special case. All this suffices, I think, to show that provided the citizens of the WOS are aware of the closeness of the alternative conceptions of the good held by their co-citizens, they could not assign with confidence a small probability to the possibility of conversion.

(b) *That the worst possible outcome would be intolerable*

Let me now move on to the second condition. An outcome is sufficiently bad as to be intolerable if we think that, having taken the risk that it will come about, we would be unable to live with the result were it to materialise. In this way intolerability is distinct from other undesirable outcomes that might come about via a calculated risk. Given the significance of the basic structure for the environment in which citizens of the WOS revise and pursue a conception of their ends, the worst possibility that could come for Betty if she voted in favour of the constitutional amendment is, it seems, very bad. If the liberties in question are to freedom of conscience and association, then we can suppose that the worst possibility amounts to having a religious conception that she is not permitted to follow or practice, the costs of which from one's own point of view will be incredibly high. If rights to bodily integrity and freedom of movement are also in question, then slavery and serfdom are possibilities, the badness of which requires no explanation. Since the burdens associated with any of these possibilities would be incredibly severe, it seems quite plausible that

the worst possibility that could come about for Betty if she voted in favour of the amendment would be intolerable.

(c) *That the status quo would be satisfactory*

Now to the third condition. Even if the worst possible outcome is intolerable, eroding or restricting the social minimum could still be rational for Betty if she places considerable value on what could potentially be gained by modifying things in her favour. However, if she found the status quo satisfactory, and placed little value on what could be gained in this way, then this would militate against the rationality of her eroding or restricting the social minimum. The most extreme possibility with respect to this condition would be if Betty were to place no value on what she could gain in this way. In this case, even the presence of a small risk of making things worse from her own point of view would render her decision vote in favour of the amendment irrational. We can therefore see that the three conditions are related in the following way: the more strongly (c) holds, the smaller the risk of an intolerable outcome needs to be to render approximating maximin rational.

Given favourable material conditions, what could potentially be guaranteed by principles of justice governing the operations of the basic structure? If Rawls's principles of justice as fairness were in place, then Betty would have the freedom to pursue her conception of the good, and sufficient material means to make use of this freedom. What she could potentially gain by voting favour of the amendment might consist in a greater share of material means for herself, or the freedom to pursue a conception of the good that involves infringing the rights of her co-citizens. These things will no doubt have some positive value from the point of view of self-interest. But for the condition to be satisfied it is not essential that they have no positive value, only that the positive value that they do have is outweighed by the risk of ending up in a less advantaged social position due to a change in one's conception of the good.

We can now examine all three conditions together and ask: could a conception of justice bring them about in the WOS? The first two conditions should be taken together. As I have said, I think there are good grounds to hold that the worst possible outcome that could come about in *Constitutional Amendment* would be intolerable for Betty. Though the particular conception of justice that is realised in her society has little impact on this: that the burdens associated with living in a social world without moral status are severe has little to do with the conception of justice that is realised in her society. The worst possibilities are as bad as they are regardless of the current situation in the WOS. By contrast, the probability that Betty assigns to the possibility that her conception of the good changes such that she comes to occupy an intolerable position can be influenced by the conception of justice that is realised in her society. It can make publicly available the knowledge that every conception of the good held in the WOS is an exercise of rationality, and that the absence of agreement among citizens is solely due to the burdens of judgment, and in doing so make the conceptions of the good held by others appear as close possibilities. However, publicising this fact appears to be the limit of what a conception of justice can do to influence citizens' probability judgments in a manner consistent with the full information condition imposed by the publicity requirement. I think, then, that the greatest possible extent to which the first two conditions can be fulfilled for my purposes hold when citizens of the WOS acknowledge the explanation that the burdens of judgment give of their continued disagreements.

Given that this is the greatest extent to which the first two conditions can hold for my purposes, the pertinent question is whether the final condition can hold to a sufficient degree. In other words, could a conception of justice bring it about that the members of the WOS are sufficiently satisfied with their lot that they would refrain from taking the risk of trying to modify the status quo to their benefit, such as by voting for the amendment in *Constitutional Amendment*? The extent to which citizens find their position satisfactory will clearly depend to a significant degree on the conception of justice that is realised in their society. If the conception does little or nothing to ensure that the prospects of the least well off social position are satisfactory, then the

citizens who occupy those positions will not be sufficiently content with their lot for it to be irrational for them to attempt to change the situation to their favour. However, if the conception of justice has as its principal aim ensuring that each of the social positions in the WOS is to some degree satisfactory for the citizens who occupy them, then we will, perhaps, be able to plausibly claim that they would be sufficiently satisfied with their lot to render the pursuit of change irrational.

6.6. Back to the Content Question

It is time to take stock, and see where the arguments given in this chapter so far leave us. In the previous chapter I defended the publicity requirement. In order to satisfy that requirement a conception of justice must be such that it would be the object of convergence in a well-ordered society. In this chapter I have focussed on the question of what, if anything, would be the object of convergence in these circumstances. Given that citizens of the WOS are fallible choosers of their conception of the good, their position in society is liable to change. They see alternative conceptions of the good as close possibilities, representing ends they themselves could come to hold, and so they see alternative social positions as positions they themselves could come to occupy. Thus if the citizens of the WOS were satisfied with their lot to a sufficient degree they would choose—even when bracketing their desire to be just for its own sake—not to decrease the prospects of alternative social positions by revoking or eroding the liberties, opportunities and material means available to them. Given all this, the question of what conception of justice satisfies the publicity requirement can be restated. It now should be understood as: what conception of justice would ensure that all social positions in the WOS are satisfactory for their occupants in this way?

This orientation of the question already tells us something quite significant. There are, after all, many conceptions of justice that do not aim to ensure that each social position is satisfactory and thus may do nothing to mitigate the possibility that some social positions are rendered

unsatisfactory. The two most established examples of such conceptions are utilitarian and Lockean views. Utilitarianism is often chastised on the basis that it would in principle require making the lives of some intolerable if doing so would bring about a greater overall balance of happiness. A route of reply to objections of this kind is to argue that, at least in ordinary circumstances, the kinds of cases the objector has in mind such as slavery or torture or killing the innocent, would not genuinely maximise utility.[270] Whatever merits this reply has to that objection, the circumstances in which utilitarianism would permit making a social position un*satisfactory* are sufficiently commonplace that that it would be hard to deny that they arise in many ordinary circumstances. Not being oriented to ensure such satisfactoriness across the board, utilitarianism will countenance various measures that render the prospects of some social positions less than satisfactory whenever doing so will generate an aggregate benefit. Therefore, utilitarianism is by design deeply unsuited to satisfying the publicity requirement.

The Lockean view, as represented by Robert Nozick's *Anarchy, State, and Utopia*, holds that whatever distribution of goods emerges from an initially just distribution of holdings via a series of just transactions is itself just.[271] Since this view guarantees the same negative rights to all, it does not face the problem faced by utilitarianism: that a social position may be made unsatisfactory by an absence of guaranteed rights and liberties. However, it offers no guarantee that citizens will end up with any level of material means at all. If the ideal process of transactions renders some destitute, then though they have their liberties, they may have little else. Without material means, negative liberties to pursue one's conception of the good are of little value from the self-interested point of view. And the position of being left with no means at all to pursue one's conception of the good would unsatisfactory: one would seek to change the status quo in one's favour if one could. Therefore, like utilitarianism, this view is deeply unsuited to ensure the satisfactoriness of social positions, and thus to satisfy the publicity requirement.

---

[270] R.M. Hare, *Moral Thinking: Its Levels, Method and Point*, (New York, NY: Oxford University Press, 1981), 130-146.
[271] (New York, NY: Basic Books, 1974), chapter 7.

No doubt there are many other conceptions of justice that would also fail to ensure that each social position was satisfactory, though utilitarianism and the Lockean view are two of the most discussed candidates. However, there will still be a number of possible conceptions of justice left on the table as potential candidates to satisfy the requirement. Conceptions that aim to secure a social minimum for all citizens, specified in terms of liberties, opportunities and material means, will be potential candidates. Provided that, in specifying this minimum, they aim to ensure that social positions are sufficiently satisfactory that their occupants would not want to change to the status quo to their benefit, given the extent to which I have said the other two conditions hold.

At this stage I want to make clear that I will not be arguing that any particular conception of justice is such that it would best satisfy the publicity requirement. I only want to show that some conceptions are ruled out while others remain potential candidates. With respect to the content question, the fact that the conception of justice that would satisfy the publicity requirement is necessarily one that renders all social positions satisfactory suffices to show that it is not radically extensionally inadequate with respect to considered moral judgments. A view would certainly be inadequate in this way if it failed to account for a fixed point in our moral thinking, such as the wrongness of slavery or genocide or similar cases. But if a conception of justice ensures that each social position is satisfactory, then I think this rules out the possibility that it fails to account for a moral fixed point, since such fixed points are given by the wrongness of actions that make some social positions intolerable or worse. How about other moral judgments that do not constitute moral fixed points, but we are nonetheless are strongly inclined to think are sound? A conception of justice that ensures that each social position is satisfactory could certainly fail to account for some such judgments, especially those that point toward utilitarian on Lockean principles. However, a failure to account for any of these weaker judgments could not amount to anything close to a *reductio ad absurdum*, and so I think we should say that, provided are confident in the premises of the argument I have offered, these judgments ought to be rejected.

7. CONCLUDING REMARKS

7.1. A Recap

A number of contemporary political philosophers have endorsed the reasonable agreement principle as a constraint on the exercise of political power. This principle holds that the exercise of political power must be justifiable to all reasonable citizens. Reasonable citizens are an idealised consistency; though as we saw in part I of this thesis, there is little agreement on how exactly they ought to be idealised. In order to defend the reasonable agreement principle, we must defend an account of what this idealisation involves. There are a number of ways in which this idealisation question might be answered, and so in order to defend a particular way of answering it, we need a rationale. Why should the exercise of political power be constrained by what reasonable citizens can be expected to accept? Or, more precisely, why should the exercise of political power be constrained by what reasonable citizens, *idealised in a particular way*, could be expected to accept? This, I argued in chapter 2, is the central question a defence of the reasonable agreement principle must answer.

In part I, I considered a variety of ways of idealising the constituency of reasonable citizens, and a number of rationales for this idealisation. Proponents of the reasonable agreement principle have appealed to: respect, civic friendship, the presuppositions of our moral attitudes, as well as stability in a number of guises, as ways of defending their particular way of idealising the constituency to whom they take acceptability to be required. Ultimately, the message of part I was that none of these arguments succeeds in giving a fully compelling answer to the question of why we ought to accept the reasonable agreement principle, at least as they are presently formulated.

In part II, I began to explore what I argued is a more promising route for a defence of the principle. Rawls's commitment to the reasonable agreement principle seems to be clearly related to the aim of showing that a society well-ordered by his conception of justice could be stable in the long run without citizens needing to be forced to comply with it. And Rawls's defenders often

appeal at some point to this aim in attempting to answer the question of why we ought to accept the principle. But as I have argued, along with others, it is far from clear why this special kind of stability is a property that a conception of justice must have. If an account could be given of why principles of justice must be stable in this way, then we would have the beginnings of a defence of a Rawlsian version of the reasonable agreement principle.

The significance of this stability requirement can be explained, I argued, by appeal to a conception of the circumstances in which our moral competence is operating free of distortions. To show that a conception of justice satisfies this stability requirement, it must be shown that citizens of a well-ordered society would converge on having a desire to act on its principles that they take to be decisive. The well-ordered society is such that each citizen has grown up under the conception of justice in question, is in conditions of full information, and is free to choose not to act on the principles of justice. These conditions, I believe, represent the conditions in which our moral competence operates free of distortions. To ask whether we would affirm a particular conception of justice in these conditions is to ask whether we would maintain our commitment to it when we abstracted away from the various distorting features of our contingent circumstances. In this way, asking whether a conception of justice would satisfy the stability requirement, and thus the reasonable agreement principle, is simply an extension of an ordinary component of moral reflection.

Of course this answer to the question of why we ought to accept the reasonable agreement principle will be unpersuasive if the version of the principle that it grounds would deliver unpalatable conclusions. If no conception of justice could satisfy the principle, or if the only conception that could satisfy it failed to account for some moral fixed points, then this would call into question the claim that the circumstances of the well-ordered society are genuinely those in which our moral competence operates free of distortions. However, I have argued that the principle is not defective in this way. By appealing to thought that citizens of the WOS would see themselves as fallible choosers of their ends, I argued that provided certain conditions hold in their

society, they would have a decisive desire to act on principles that express their fallibility. To express our fallibility is to abstract away from our conception of the good and other contingent features of our circumstances when making political decisions, in acknowledgement of the fact that in the future we may modify or reject this conception what is of value. The desire to express one's fallibility is thus a version of the Kantian desire to achieve autonomy and avoid heteronomy, at least as Rawls reads Kant.[272]

Acknowledging that our view may change in this way is to accept that our social position may also change, perhaps dramatically. In ordinary circumstances, this possibility may well have little pull on our practical reasoning. However, I have argued that in particular circumstances this desire would move citizens of the well-ordered society decisively. Those circumstances are simply the circumstances in which maximising the fulfilment of one's interests would amount to make one's reasoning approximate the maximin decision rule, taking the course of action that would render the least bad possibility as good as possible. A conception of justice can bring about these conditions in a well-ordered society principally by ensuring that every social position is satisfactory from the point of view of its occupant. If each citizen is sufficiently satisfied with her lot under a particular conception of justice, she will not want to act against its principles where that involves a risk of setting back the interests of her future self as the occupant of a different social position, or of others about whom she cares or may come to care.

Understood in this way, the reasonable agreement principle directs us toward a conception of justice that will ensure that every social position in the well-ordered society is satisfactory from the point of view of its occupant. Though I have not argued that a particular conception of justice best achieves this aim, I have argued that those which do nothing to ensure the least well off social position is satisfactory will thus clearly fail to satisfy the principle. Given this implication, I believe that the principle does not fail to account for any fixed points in our moral thinking. Since it is not

[272] *A Theory of Justice*, 222.

defective in this way, and it has what I think is a powerful rationale for the idealisation it employs, I believe that this version of the reasonable agreement principle is plausible and compelling.

7.2. Ideals of Justice and the Good

There are two broader questions that might be asked about the implications of the argument for the reasonable agreement principle that I have presented. One concerning the ultimate status of comprehensive doctrines under the principle, and the other concerning the status of ideals of justice. In order to consider these questions, it will be useful to have the two central arguments of part II of the thesis before us. According to the argument from convergence:

(1) We possess a moral competence.

(2) If we were in the circumstances in which our moral competence operates free of distortion, our normative outlooks would converge.

(3) The principles on which we would converge in these circumstances are correct.

(4) Our moral competence operates free of distortion in circumstances C.

(5) In circumstances C we would converge on conception of justice J.

(6) Therefore, conception of justice J is correct.

The conditions in which our moral competence operates free of distortions are those of a well-ordered society in circumstances of full publicity. Therefore, discerning what conception of justice we would converge on in circumstances C requires us to work out what conception of justice would be such that citizens would have a decisive desire to act on its principles in conditions of full publicity. According to the argument from the mutability of social positions:

(1) Citizens of the well-ordered society accept that they are fallible choosers of their particular ends, and thus that they may come to modify or abandon them.

(2) How well particular social positions in the well-ordered society fare depends on the ends of their occupants.

(3) It is possible that a citizen will modify her ends such that she comes to occupy a less advantaged social position.

(4) The circumstances of the well-ordered society are such that the following conditions hold to a sufficient degree to make approximating the maximin rule rational.

(a) citizens have no reliable basis for estimating the probability that a change in their ends will lead them to occupy a less advantaged social position.

(b) the worst possible outcome, if the status quo is modified and a citizen's ends change, would be intolerable.

(c) citizens find the status quo satisfactory, and care little for what they might gain by changing it to their personal advantage.

(5) Therefore, a citizen of the well-ordered would not want to act on principles that aim to change the status quo to her personal advantage.

If a conception of justice brings about conditions (4)(a)-(c) in the well-ordered society, then citizens of that society will choose to act on, rather than against, its principles, and that conception will thereby be the object of convergence in the circumstances in which our moral competence operates free of distortions.

The first question that arises from taking these arguments in together is this. It was generally taken to be an advantage of Rawls's political liberalism that a citizen could accept his conception of justice while retaining a wholehearted commitment to their comprehensive doctrine. The conception of justice puts itself forward as 'reasonable', rather than 'true' for precisely the reason of wanting to avoid conflicts with comprehensive doctrines that may reserve the concept of truth for a special set of claims within that doctrine, and thus deny that these principles of justice could be true.[273] Citizens are meant to be able to incorporate the conception of justice for the basic structure of society into their broader worldview, which may contain claims about metaphysics,

[273] *Political Liberalism*, 94.

epistemology, and about what values ought to govern other areas of human life and its relationship to the natural world.[274]

From the two arguments I have given, it may seem that it is not possible to be fully committed one's comprehensive doctrine while accepting this case for the reasonable agreement principle (and thus for a conception of justice that ensures that each social position is satisfactory from the point of view of its occupant) that I have made. This is so because the claim that if we were in circumstances where our moral competence was operating free of distortions our normative outlooks would converge is quite general. It is not only that we would converge on the correct conception of justice, but also that we would converge on a whole set of normative beliefs. Since all I have argued would be the object of convergence in these circumstances is a conception of justice with a particular content, this may seem to be to deny that any other normative beliefs are correct. If the arguments, taken together, imply this, then citizens will not be able to accept them while also seeing their broader worldview as true. The view would therefore not have the attractive feature of allowing citizens to have a wholehearted commitment to their comprehensive doctrine alongside it.

I do not think that the arguments have this implication however. We have no easy or direct access to what we would affirm in circumstances where our moral or other normative competences operate free of distortions. The argument that I presented in chapter 6 offers an answer to the sceptic who denies that anything would be object of convergence in these circumstances, or that only an unacceptable view would be. But to accept that argument is not necessarily to hold that no other claims would also be the object of convergence in these circumstances. The argument is silent on whether or not they would be. If a citizen thinks they have done all they can to abstract away from the features of their contingent circumstances that might distort their judgment in coming to their broader worldview, then they may conjecture that others in undistorted circumstances would share this view. Moreover, though I have argued that the claims on which we

[274] See Rawls's comments on his conception of justice as module that can fit into various reasonable comprehensive doctrines, e.g. *Political Liberalism*, 12.

would converge in these circumstances are correct, I have not claimed that there are no correct claims which would not be the object of convergence. In these ways, the arguments I have offered for the reasonable agreement principle need not imply that an individual citizen's broader worldview is false.

The second question that arises is as follows. I have argued that the version of the reasonable agreement principle I defend rules out certain conceptions of justice as plausible candidates to be the object of convergence. Namely, those that do nothing to ensure that all social positions are satisfactory from the point of view of their occupants such as utilitarian and Lockean views. However, this leaves on the table a number of conceptions of justice that are not flawed in this way. This raises the question: what is the status of these conceptions? Though I have not argued that a particular conception of justice is best, I think the view that I have defended allows us to see what considerations are relevant to this question, and what modes of argumentation would support some conceptions of justice over others. I offer two examples of kinds of argument that this version of the reasonable agreement principle makes especially pertinent. The first is Rawls's argument that there is no acceptable way to select a satisfactory social minimum and adjust it to changing circumstances that does not depend on the difference principle.[275] Since the reasonable agreement principle requires such a satisfactory minimum, if it could be shown that there was no other plausible way specifying the idea then that would count strongly in favour of the difference principle. The second is the argument, due in one form to Joshua Cohen, that once we accept the goal of ensuring that every social position in the well-ordered society is satisfactory from the point of view of its occupant, there is a strong intuitive case for the difference principle. For if we have done all we can to ensure that those in the least well-off position are satisfied with their lot, then we will have done all we can to ensure each social position is satisfactory.[276] In both of these arguments the difference principle's focus on the least well-off position is defended as a

[275] *A Theory of Justice*, 278-279. See also the critical discussion in Jeremy Waldron, "John Rawls and the Social Minimum," *Journal of Applied Philosophy*, 3(1), 1986, 21-33.
[276] Joshua Cohen, "Democratic Equality," *Ethics*, 729.

uniquely successful strategy for ensuring that all social positions are satisfactory. I will not take a stand on whether either argument succeeds, but it should be clear that they take a renewed relevance highly relevant to the question of which conception of justice best satisfies the desiderata defended here.

Though there will no doubt always be disagreement about which conception of justice is best, I hope that the account I have offered here presents a fruitful way of framing such disputes. If the greatest theoretical virtue a conception of justice can have is being the one that we would affirm when our powers of moral judgment operate free of distortions, then we must, I believe, affirm a conception that ensures that the position of every citizen is satisfactory.

Bibliography

Baier, Kurt, *The Moral Point of View*, (Ithaca, NY: Cornell University Press, 1958).

Barry, Brian, *A Treatise on Social Justice Volume II: Justice as Impartiality*, (Oxford: Clarendon Press, 1995).
———"John Rawls and the Search for Stability," *Ethics*, 105(4), 1995, 874-915.

Billingham, Paul, "Does political community require public reason? On Lister's defence of political liberalism," *Politics, Philosophy & Economics*, 15(1), 2016, 20-41.
———"Liberal Perfectionism and Quong's Internal Conception of Political Liberalism," *Social Theory and Practice,* 43(1), 2017, 79-106.

Buchanan, Allen, "Revisability and Rational Choice," *Canadian Journal of Philosophy*, (5)3, 1975, 395-408.

Caney, Simon, "Anti-Perfectionism and Rawlsian Liberalism," *Political Studies*, 43(2), 1995, 248-264.
———"Impartiality and Liberal Neutrality," *Utilitas*, 8(3), 1996, 273-293.

Christensen, David, "Epistemology of Disagreement: The Good News," *Philosophical Review,* 166(2), 2007, 187-21.

Cohen, G.A., *Rescuing Justice and Equality*, (Cambridge, MA: Harvard University Press, 2008).

Cohen, Joshua, "Democratic Equality," *Ethics*, 99(4), 1989, 727-751.

Crisp, Roger, *Reasons and the Good*, (Oxford: Oxford University Press, 2006).

Darwall, Stephen, Gibbard, Allan, and Railton, Peter, "Toward *Fin de siècle* Ethics: Some Trends," *The Philosophical Review*, 101(1), 1992, 115-189.

Darwall, Stephen, *The Second-person Standpoint*: *Morality, Respect and Accountability*, (Cambridge, MA: Harvard University Press, 2006).

Ebbels-Duggan, Kyla, "The Beginning of Community: Politics in the Face of Disagreement," *The Philosophical Quarterly*, 60(238), 2010, 50-71.

Enoch, David, "Against Public Reason," *Oxford Studies in Political Philosophy: Volume 1*, (New York, NY: Oxford University Press, 2015), 117-130.
———"Why Idealize?" *Ethics*, 115(4), 2005, 759-787.

Fishkin, James S., "Review: *Patterns of Moral Complexity*," *Political Theory*, 17(1), 1989, 153-156.

Frankfurt, Harry, "Alternative Possibilities and Moral Responsibility," *The Journal of Philosophy* 66(23), 1969, 829-839.
———"Freedom of the Will and the Concept of a Person," *The Journal of Philosophy*, 68(1), 1971, 5-20.

Freeman, Samuel, "Congruence and the Good of Justice," *A Cambridge Companion to Rawls*, edited by Samuel Freeman, (Cambridge: Cambridge University Press, 2003), 277-316.

Gaus, Gerald, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, (New York, NY: Cambridge University Press, 2011).
——— *Justificatory Liberalism: An Essay on Epistemology and Political* Theory, (New York, NY: Oxford University Press, 1996).
——— "The Rational, the Reasonable and Justification," *The Journal of Political Philosophy*, 3(3), 1995, 234-258.
———*Value and Justification: The Foundations of Liberal Theory*, (New York, NY: Cambridge University Press, 1990).
——— "The Turn to a Political Liberalism," in *A Companion to Rawls*, edited by Jon Mandle and David A. Reidy, (Oxford: Wiley-Blackwell, 2014), 235-250.

Gauthier, David, *Morals by Agreement*, (Oxford: Oxford University Press, 1999).

Gibbard, Allan, *Wise Choices, Apt Feelings: A Theory of Normative Judgement*, (Cambridge, MA: Harvard University Press, 1993).

Habermas, Jürgen, *Moral Consciousness and Communicative Action*, (Cambridge, MA: MIT Press, 1990).

Hare, R.M., *Moral Thinking: Its Levels, Method and Point*, (New York, NY: Oxford University Press, 1981).

Harsanyi, John, "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *The American Political Science Review*, 69(2), 1975, 594-606.

Hooker, Brad, *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*, (Oxford: Oxford University Press, 2000).
———"Publicity in Morality: A Reply to Katarzyna de Lazari-Radek and Peter Singer," *Ratio*, XXIII(1), 2010, 111-117.

Jensen, Mark, "The Limits of Practical Possibility," *The Journal of Political Philosophy*, 17(2), 2009, 168-184.

Joyce, Richard, *The Evolution of Morality*, (Cambridge, MA: MIT Press, 2006).

Kahane, Guy, "Evolutionary Debunking Arguments," *Nous*, 45(1), 2011, 103-125.

Kant, Immanuel, *Groundwork of the Metaphysics of Morals*, Edited by Mary Gregor and Jens Timmermann, (New York, NY: Cambridge University Press, 2012[1785]).
——— *Practical Philosophy*, Translated by Mary J. Gregor, (Cambridge: Cambridge University Press, 1996).

Kelly, Thomas, "The Epistemic Significance of Disagreement," *Oxford Studies in Epistemology: Volume 1*, edited by Tamar Gendler Szabo and John Hawthorne, (Oxford: Oxford University Press, 2005), 167-196.

Kramer, Matthew, *Moral Realism as a Moral Doctrine*, (Oxford: Wiley-Blackwell, 2009).

Krasnoff, Larry, "Consensus, Stability, and Normativity," *The Journal of Philosophy*, 95(6), 1998, 269-292.

Laden, Anthony, "Games, Fairness, and Rawls's *A Theory of Justice*," *Philosophy & Public Affairs*, 20(3), 1991, 189-222.

Larmore, Charles, *Patterns of Moral Complexity*, (New York, NY: Cambridge University Press, 1987).
———"Political Liberalism," *Political Theory*, 18(3), 1990, 339-360.

Leland, R.J., and van Wietmarschen, Han, "Political Liberalism and Political Community," *Journal of Moral Philosophy*, 14(2), 2017, 142-167.

Lenman, James, and Shemmer, Yonotan, "Introduction," in *Constructivism in Practical Philosophy*, edited by James Lenman and Yonatan Shemmer, (Oxford: Oxford University Press, 2012).

Lewis, David, "Dispositional Theories of Value," *Proceedings of the Aristotelian Society*, Supplementary Volume 63, 1989, 133-137.

Lister, Andrew, *Public Reason and Political Community*, (London: Bloomsbury, 2013).

Macedo, Stephen, *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism*, (New York, NY: Clarendon Press, 1991).

Mang, Franz Fan-lun, "Liberal Neutrality and Moderate Perfectionism," *Res Publica*, 19(4), 2013, 297-315.

Nagel, Thomas, "Moral Conflict and Political Legitimacy," *Philosophy & Public Affairs*, 16(3) 1987, 215-240.

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

O'Neill, Martin, *Freedom, Fairness and Responsibility*, (Harvard PhD Dissertation, 2008).

O'Neill, Onora, "Constructivism vs. Contractualism," *Ratio*, XVI(4), 2003, 319-331.

Parfit, Derek, *On What Matters: Volume One*, (New York, NY: Oxford University Press, 2011).
———*On What Matters: Volume Two*, (New York, NY: Oxford University Press, 2011).

Pew Research Group, "Faith in Flux: Changes in Religious Affiliation in the U.S." *The Pew Forum on Religion and Public Life*, 2009. Retrieved on 6 September 2017 from http://www.pewforum.org/2009/04/27/faith-in-flux/

Plato, *The Republic*, translated by G.M.A. Grube, revised by C.D.C. Reeve, (Indianapolis, IN: Hackett, 1992), book I.

Quong, Jonathan, *Liberalism without Perfection*, (New York, NY: Oxford University Press, 2011).
———"What is the point of public reason?" *Philosophical Studies*, 170(3), 2014, 545-553.

Rawls, John, *A Theory of Justice: Revised Edition*, (Cambridge, MA: The Belknap Press of Harvard University Press, 1990).
———*Political Liberalism: Expanded Edition*, (New York, NY: Columbia University Press, 2005).
———*Collected Papers*, (Cambridge, MA: Harvard University Press, 1999).
———*Justice as Fairness: A Restatement*, (Cambridge, MA: The Belknap Press of Harvard University Press, 2001).

Raz, Joseph, "Facing Diversity: The Case of Epistemic Abstinence," *Philosophy & Public Affairs*, 19(1), 1990, 3-46.

Rousseau, Jean-Jacques, *The Social Contract and the Discourses*, translated by G.D.H. Cole, (London: Everyman, 1993[1762]).

Sandel, Michael, *Liberalism and the Limits of Justice: Second Edition*, (Cambridge, Cambridge University Press, 1998).

Scanlon, T.M., *What We Owe to Each Other*, (Cambridge, MA: The Belknap Press of Harvard University Press, 1998).
———"Rawls on Justification," in *The Cambridge Companion to Rawls*, edited by Samuel Freeman, (Cambridge: Cambridge University Press, 2002), 139-167.

Schafer-Landau, Russ, *Moral Realism: A Defense*, (Oxford: Oxford University Press, 2003).

Schroeder, Mark, "Having Reasons," *Philosophical Studies*, 139(1), 2008, 57-71.

Seglow, Jonathan, "Neutrality and Equal Respect: On Charles Larmore's Theory of Political Liberalism," *The Journal of Value Inquiry*, 37(1), 2003, 83-96.

Sidgwick, Henry, *The Methods of Ethics*, (Indianapolis, IN: Hackett, 1981[1907]).

Singer, Peter, and de Lazari-Radek, Katarzyna, "The Objectivity of Ethics and the Unity of Practical Reason," *Ethics*, 123(1), 2012, 9-31.

Smith, Michael, *The Moral Problem*, (Oxford: Blackwell Publishing, 1994).

Sobel, David, "Do the desires of rational agents converge?" *Analysis* 59(3), 1999, 137-147.

Strawson, P.F., *Freedom and Resentment and Other Essays*, (Oxford: Routledge, 2008[1962]).

Street, Sharon, "What is Constructivism in Ethics and Metaethics?" *Philosophy Compass*, 5(5), 2010, 363-384.
———"A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies*, 127(1), 2006, 109-166.

Taylor, Robert, "Rawls's Defense of the Priority of Liberty: A Kantian Reconstruction," *Philosophy & Public Affairs*, 31(3), 2003, 246-271.

Vallier, Kevin, "Convergence and Consensus in Public Reason," *Public Affairs Quarterly*, 25(4), 2011, 261-280.
———*Liberal Politics and Public Faith: Beyond Separation*, (New York, NY: Routledge, 2014).

Vavova, Katia, "Debunking Evolutionary Debunking," *Oxford Studies in Metaethics: Volume 9*, 2014, 76-101.

Waldron, Jeremy, "Theoretical Foundations of Liberalism," *The Philosophical Quarterly*, 37(147), 1987, 127-150.
———*Law and Disagreement*, (Oxford: Oxford University Press, 1999).
———"John Rawls and the Social Minimum," *Journal of Applied Philosophy*, 3(1), 1986, 21-33.
———"Hobbes and the Principle of Publicity," *Pacific Philosophical Quarterly*, 82(3-4), 2001, 447-474.

Wallace, R.J., *Responsibility and the Moral Sentiments*, (Cambridge, MA: Harvard University Press, 1996).

Weber, Max, *From Max Weber: Essays in Sociology*, (London: Routledge, 1991).

Weithman, Paul, *Why Political Liberalism? On John Rawls's Political Turn*, (New York, NY: Oxford University Press, 2013).

Wenar, Leif, "Political Liberalism: An Internal Critique," *Ethics*, 105(1), 1995, 32-62.

Williams, Bernard, *Moral Luck: Philosophical Papers 1973-1980*, (Cambridge: Cambridge University Press, 1981).
———*In the Beginning was the Deed: Realism and Moralism in Political Argument*, (Princeton, NJ: Princeton University Press, 2005).
———"A Critique of Utilitarianism," in *Utilitarianism: For and Against*, edited by J.J.C. Smart and Bernard Williams, (Cambridge: Cambridge University Press, 1973), 77-150.