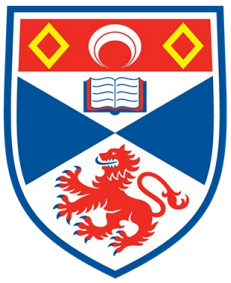
Disagreement, Concepts and Convergence: A New Theory of Political Realist Legitimacy

Saranga Sudarshan



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of
Doctor of Philosophy (PhD)
at the University of St Andrews

April 2020

Candidate's declaration

I, Saranga Sudarshan, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 78,768 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2016.

I confirm that no funding was received for this work.

Date 16/07/2020

Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date 1707/2020

Signature of supervisor

Date 16/07/2020

Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Saranga Sudarshan, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 16/07/2020

Signature of candidate

Date 16/07/2020

Signature of supervisor

Date 16/07/2020

Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Saranga Sudarshan, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 16/07/2020

Signature of candidate

Disagreement, Concepts and Convergence: A New Theory of Political Realist Legitimacy

Saranga Sudarshan

University of St Andrews and University of Stirling

2020

Typeset with LaTeX 2ε using Ubuntu 20.04
Saranga Sudarshan 2020

Abstract

This thesis argues for a novel conception of political realism as a theory of political legitimacy: the Dual Convergent Conception. The thesis is framed by the thought that one way of theorising about political legitimacy involves working out how reasonable people can achieve a stable political order so that, despite their profound moral differences, they may live together governed by principles they have sufficient moral reason to affirm from within their own point of view. I argue that this ultimately involves making a special sort of argument that takes reasonable disagreement about justice seriously: a Disagreement to Legitimacy argument. This is an argument with two parts. The first part involves finding the best explanation of reasonable disagreement about justice. After arguing against all extant explanations of reasonable disagreement, I develop a novel explanation: Diverse Packages Theory. This explanation makes use of the idea of metalinguistic negotiation and empirical work in developmental psychology on concepts, to argue that the best explanation of reasonable disagreement about justice is that reasonable people possess and use diverse concepts and conceptions of justice. The second part of the Disagreement to Legitimacy argument involves proposing, on the basis of Diverse Packages Theory's explanation, how all reasonable people can have sufficient moral reason to coordinate and continue coordinating over time on coercive principles or rules that order society's basic institutions. I then argue that extant conceptions of political liberalism and political realism cannot show how reasonable people can achieve this. I then argue that by combining certain elements of the political liberal view of convergent agreements, and the political realist focus on a contextually situated acceptance of coercively enforced political principles, the Dual Convergent Conception of political realism can show how reasonable people can achieve a stable political order.

Acknowledgements

I would first like to thank my supervisors over the years in the St Andrews and Stirling Graduate Program. Ben Sachs, Derek Ball, and Simon Hope have provided the rigorous supervision and encouragement that has made this thesis possible. In addition, I would like to thank Theron Pummer and Kent Hurtig who as my annual reviewers provided the critical guidance that kept me on track. I would like to thank Kevin Scharp and Patrick Greenough for running the Arche Conceptual Engineering Seminar and providing guidance which has helped me immeasurably. I would like to thank Gerald Gaus and the graduate community at the University of Arizona from whom I learnt a great deal about doing political philosophy.

My second round of thanks goes to the following graduate students in the St Andrews and Stirling Graduate Program that have helped me think through issues, provided comments on early work and helped proofread: Lisa Bastian, Joseph Bowen, Miguel Egler, Claire Field, Mirela Fus, Jakob Hinze, Lara Jost, Ethan Landes, Lorenzo Lazzarini, Stefano Lo Re (who read almost the entire thesis), Thomas McKenna, Colin McLean, Anh Quan Nguyen, Quentin Pharr, Lewis Ross, Alessandro Rossi, Janis Schaab, Philipp Schönegger, Lucas Sierra Velez, Joseph Slater, Ravi Thakral, and Clotilde Torregrossa. More than any individual contribution the graduate community at St Andrews encouraged an atmosphere of serious academic research, where the norm of debate was 'to give no quarter and ask none'. I hope that norm continues.

A PhD is a long journey that starts before a graduate program. To that end, I would like to thank those at the University of Sydney who provided my early education in philosophy, and especially the following people: Thomas Besch who introduced me to Rawls and political philosophy generally, John Grumley who introduced me to the diversity of political philosophy, and David Braddon-Mitchell for the "Thought and Hyperintensionality" graduate seminar in 2014 which seeded many of the ideas in this thesis.

And of course lastly, I would like to thank my parents and my little sister, who encouraged and supported my love of philosophy.

Contents

1 From Disagreement to Political Legitimacy 5
- 1 Introduction . . . 5
- 2 Theorising about Political Legitimacy . . . 7
- 3 Reasonable Disagreement about Justice as an Explanandum . . . 23
- 4 Road Ahead . . . 29

2 Explaining Reasonable Disagreement about Justice 31
- 1 Introduction . . . 31
- 2 Two Enrichments . . . 32
- 3 Imperfection Family . . . 36
- 4 Historical-Psychological Family . . . 49
- 5 Conceptual Family . . . 58
- 6 Conclusion . . . 77

3 Diverse Packages Theory 79
- 1 Introduction . . . 79
- 2 Diverse Packages Theory . . . 80
- 3 Applying the Model . . . 93
- 4 Comparative Advantages . . . 105
- 5 Objections . . . 106
- 6 Conclusion . . . 112

4 The Instability of Political Liberalism 115
- 1 Introduction . . . 115
- 2 Consensus Conception . . . 117
- 3 Convergence Conception . . . 128
- 4 Conclusion . . . 155

5 The Instability of Political Realism 157
1 Introduction 157
2 Non-Domination Conception 159
3 Restrained Domination Conception 168
4 Conclusion 178

6 Sketch of a New Political Realism 179
1 Introduction 179
2 Dual Convergent Conception 181
3 Modus Vivendi Objection 196
4 Conceptual Integrity Objection 197
5 Conclusion 198

Chapter 1

From Disagreement to Political Legitimacy

1 Introduction

Let me start this thesis where I hope to end it. At the end of *A Theory of Justice*, in describing the kind of perspective that he has tried to adopt in his arguments and which give them their normative force, Rawls (1999: 514) says:

> The perspective of eternity is not a perspective from a certain place beyond the world, nor the point of view of a transcendent being; rather it is a certain form of thought and feeling that rational persons can adopt within the world. And having done so, they can, whatever their generation, bring together into one scheme all individual perspectives and arrive together at regulative principles that can be affirmed by everyone as he lives by them, each from his own standpoint. Purity of heart, if one could attain it, would be to see clearly and to act with grace and self-command from this point of view.

What Rawls sums up there and continues in *Political Liberalism* is doing political philosophy in a way that tells us how we ought to act from the "perspective of eternity". In *A Theory of Justice* this perspective involves bringing together people pursuing diverse conceptions of the good life (or as he calls them "rational plans of life") and crafting a conception of justice that allows them to live together in a stable political order. In *Political Liberalism*, or at least in one part of the book's project, this perspective is expanded to diverse conceptions of justice itself (or as he calls them "comprehensive doctrines") and crafting the grounds of a *legitimate* conception of justice. That is where this thesis begins, with the perspective of eternity involving reasonable

disagreement about justice itself. This is because of the simple fact that everywhere we look in society reasonable people disagree not merely about what is good for them, but also about what is ultimately good for other people, for society, and how this can be rightfully achieved. Or in other words, they disagree about justice itself. This involves doing political philosophy in a way that takes seriously, as Charles Larmore (2013: 295) says, "the modern realization that reasonable people tend naturally to disagree about fundamental aspects of the right and the good".

As it was for Rawls, when we move to taking reasonable disagreement about justice seriously, we move from theorising about justice alone to theorising about political legitimacy. The goal being to show how we can achieve a stable political order so that people with profound moral differences may live together governed by principles they can all affirm from within their own point of view and not, as Rawls (2005: lx) says, "be governed by power and coercion alone". That is what this thesis is about. It is devoted to arguing for a theory of political legitimacy that takes reasonable disagreement about justice seriously. This is a theory that tells us which political principles or rules that are coercively enforced reasonable people have sufficient moral reason to coordinate on.[1] This will tell us how reasonable people can achieve a stable political order.

For Rawls and his followers, the theory of political legitimacy that achieves all this is a conception of political liberalism. In short, that a political principle or rule is legitimate if all reasonable people conclusively justify *endorsing* it. This sort of free endorsement is what, according to political liberals, can achieve a stable political order where people with profound moral differences can live together by political principles or rules they call affirm from within their own point of view.

In this thesis I argue political liberalism is wrong about what constitutes a stable political order. I argue that we ought to embrace a conception of political realism. In short, that a political principle is legitimate if all reasonable people conclusively justify *accepting* it in a given context. This sort of contextually situated acceptance is what, according to political realists, can achieve a stable political order.

But, I will argue that extant conceptions of political realism, just like political liberalism, also cannot achieve a stable political order. Given that, I propose a novel conception of political realism: the Dual Convergent Conception. The key idea being that political realism involves a state of affairs of 'ordered moral warfare'. This is a state of affairs where all reasonable people can accept a political principle or rule and yet be

[1] Throughout this thesis I will take the object of political legitimacy to be a "political principle or rule" because it is a neutral way of describing what various theorists seek to legitimate. Whether those principles and rules are part of a 'theory' or merely laws will be left to competing theorists to specify.

able to advocate for their ideal political principles or rules and tolerate others doing so as well. And, crucially that this is not a modus vivendi where two disagreeing parties refrain from coercing the other for prudential reasons.

The main argument for the Dual Convergent Conception will be that it, as opposed to its competitors, actually takes reasonable disagreement about justice seriously. I will argue that it responds better than its competitors to what the best explanation of reasonable disagreement about justice tells us actually causes such disagreement. This will show, I submit, that the Dual Convergent Conception does better than its competitors, at showing how reasonable people, despite their reasonable disagreements, can achieve a stable political order. This is because it is better than its competitors at showing which political principles or rules that are coercively enforced all reasonable people have sufficient moral reason to coordinate on. As such, much of this thesis will involve showing how, in light of what I will argue is the best explanation of reasonable disagreement about justice, theories of political legitimacy either succeed or fail at dealing with what causes reasonable disagreement about justice whilst not violating some moral fixed points about the use of coercive power and the kind of stable political order we require.

Given all that, the goal of this chapter is to clear the ground for making the argument for the Dual Convergent Conception. To that end I settle the details of what it means to argue for a theory of political legitimacy that takes reasonable disagreement about justice seriously and how this sets the trajectory of the rest of the thesis. The argument of this chapter then proceeds as follows. In §2 I settle the details of what I will take political legitimacy to be and propose that theorising about political legitimacy involves making a special sort of argument: a Disagreement to Legitimacy argument. I then consider and respond to two objections to Disagreement to Legitimacy arguments. In §3 I tease out the methodological implications of theorising about political legitimacy in this way and how it will determine the trajectory of my argument for the Dual Convergent Conception. Finally in §4 I detail the road ahead and how this thesis will proceed.

2 Theorising about Political Legitimacy

In this section I lay out the methodology of this thesis. I clarify the idea of political legitimacy, describe what it means to theorise about it in terms of Disagreement to Legitimacy arguments, and why making such an argument is the core of this thesis. In §2.1 I clarify how I understand political legitimacy to involve reasonable people

having sufficient moral reason to coordinate on a political principle or rule that is coercively enforced. I argue that it involves both the justification of political authority and the use of coercive power. This is because only a theory that includes both notions can show us how to achieve a stable political order. In §2.2 I argue that we ought to conceive of theorising about political legitimacy as making Disagreement to Legitimacy arguments. This is because it clarifies the formal features of a project that takes reasonable disagreement about justice seriously and so provides a clear methodological framework in which different theories of political legitimacy can be evaluated. In §§2.3–2.4 I consider two objections that might be raised against Disagreement to Legitimacy arguments. First, I consider the objection that my characterisation of a Disagreement to Legitimacy arguments does not square with those who I claim make such arguments. Second, I consider the objection that whatever Disagreements to Legitimacy arguments do, they cannot involve arguing for a theory of political legitimacy. This is because fundamental normative principles are fact-insensitive and the whole point of a Disagreement to Legitimacy argument involves normative theories being sensitive to the fact of reasonable disagreement. I respond to both objections in turn and defend Disagreement to Legitimacy arguments as a way of theorising about political legitimacy.

2.1 What is Political Legitimacy?

As I have said, this thesis is about arguing for a theory of political legitimacy. But, before I can explain what it means to do such a thing, we need to get clear on what political legitimacy is. In short, what does it mean to say that a political principle or rule is legitimate? So far I have talked of political legitimacy in general terms. I have said that it involves reasonable people having sufficient moral reason to coordinate on a political principle or rule that is coercively enforced. Let me now be more specific about what that means.

I take "reasonable people having sufficient moral reason to coordinate on a political principle or rule that is coercively enforced" to mean the justification of political authority and the justification of coercion. An example of the first justification is the justification of principles or rules that describe what justice requires. This justification establishes the authority of these political principles or rules which means they are principles or rules that people ought to obey. An example of the second justification is the justification of the use of coercive power to enforce the principles or rules. This justification establishes the permissibility of using coercive power to enforce political principles or rules. These dual justifications are then cashed out more simply as

reasonable people having *sufficient moral reason* to coordinate on a political principle or rule that is coercively enforced. A *moral reason* here is a reason that reasonable people see as a moral reason from their point of view. These are reasons that they see as having the type of normativity that makes coercively enforced political principles or rules authoritative. A *theory* of political legitimacy then explains the existence conditions of reasonable people having sufficient moral reason. It explains how reasonable people, despite their disagreements about justice, can have sufficient moral reason to coordinate on a political principle or rule that is coercively enforced.

Now this might strike some readers as odd. If political legitimacy is about justifying the authority of a political principle or rule, why do we need any mention of coercion? Conversely, if political legitimacy is about justifying the coercion to enforce a political principle or rule, why do we need any mention of authority? To answer both questions, and to explain why I use the idea of political legitimacy above, let me clear the ground by considering two popular ways of thinking about political legitimacy.

On one way of thinking about political legitimacy it involves justifying *political authority* (Raz 2006, 1986; Bird 2014; Simmons 2001; Edmundson 1998; Green 1988). It involves justifying a political principle or rule about what individuals, social institutions and social systems ought to do and what they owe to each other. In short, establishing that the political principle or rule is authoritative. Here political legitimacy does not involve justifying the use of coercive power or any notion of coercion at all. It involves justifying a political principle or rule, that may or may not be enforced by coercion, as a political principle or rule that ought to be obeyed.

On this view, a theory of political legitimacy that takes reasonable disagreement about justice seriously responds to the fact that reasonable people each claim moral authority over one another. This is because they view the political principle or rule they propose as an expression of what justice requires. A theory of political legitimacy then justifies whose political principle or rule is the one that reasonable people ought to obey. This means, explaining why reasonable people have sufficient moral reason to coordinate on a particular political principle or rule over another. The underlying idea is that political legitimacy is not about justifying coercion. It is about justifying political authority.

But, on another way of thinking about political legitimacy it involves justifying the *use of coercive power* (Rawls 2005; Ripstein 2004; Buchanan 2002; Nagel 1991, 1987). Here political legitimacy here is not about establishing a political principle or rule as authoritative. Rather political legitimacy is about justifying the use of coercive power to enforce a particular political principle or rule.

On this view, a theory of political legitimacy that takes reasonable disagreement about justice seriously responds to the fact that reasonable people aim to use coercive power against each other to enforce the political principles or rules they see as expressions of what justice requires. What needs to be established is whose use of coercive power is morally permissible. The theory then explains what justifies the use of coercive power to enforce a political principle or rule. The underlying idea is that, political legitimacy is not about justifying a political principle or rule as authoritative, but about justifying the use of coercive power to enforce it.

Rather than trying to argue that one of those views of political legitimacy is the correct one, I follow Thomas Christiano (2009: 240–243) and others in taking a broader view of political legitimacy that incorporates both views.[2] The idea being that political legitimacy is about the justification of a political principle or rule as authoritative *and* the justification of the use of coercive power to enforce that political principle or rule. This amounts to the idea that political legitimacy involves justifying both the moral authority of a political principle or rule and the use of coercive power to enforce that particular political principle or rule. Both justifications together amount to reasonable people having sufficient moral reason to coordinate on a political principle or rule that is coercively enforced.

The reason for adopting this broader view of political legitimacy is that only a theory that involves this view can fulfil its function of showing reasonable people how to achieve a stable political order. To see why, consider what I said the point of this thesis was. It was to argue for a theory of political legitimacy that takes reasonable disagreement justice seriously. In such a situation reasonable people disagree about the political principles or rules that should order a society so people can pursue their conceptions of the good life and resolve claim disputes. That is after all the function of justice.[3] As such, the point of a theory of political legitimacy is to describe how reasonable people who disagree about justice can live together in a shared social world by explaining which political principles and rules that order a society can be justified to them and continued to be justified to them over time.

This means, if political legitimacy was only about justifying the moral authority

[2] See Gaus (2011b: 460–470) for another example of a theory of political legitimacy that incorporates both aspects. Strictly speaking Christiano lists three aspects of political legitimacy with the creation of duties owed to the state being separate from the creation of duties owed between individuals. For the purposes of this thesis nothing hangs on this distinction. A conception of justice and even specific laws can do both.

[3] Rawls (1999: 5–6, 9) is perhaps the most prominent contemporary example of this way of viewing justice, but many theorists – too many to mention here – have viewed justice in the same way. See Schoelandt (2020) for a detailed overview.

of a political principle or rule, the theory would not be able to say anything about the moral permissibility of the coercion needed to enforce the political principle or rule so it actually fulfils its function of ordering a society. It would not actually be able to show why it is morally permissible to use coercive power against reasonable people acting in ways they have insufficient moral reason to act, or against unreasonable people acting for non-moral reasons. It would merely explain why reasonable people ought to coordinate on a particular political principle or rule and not another. Without saying anything about the moral permissibility of coercively enforcing a political principle or rule, the theory would be unable to show how to achieve a stable political order amongst the unreasonable or immoral reasonable people who do not obey the political principle of rule.

On the other hand, if political legitimacy was only about justifying the use of coercive power that enforces a political principle or rule, the theory would say nothing about why that political principle or rule in particular ought to be enforced. This would mean a theory of political legitimacy would not actually be able to show why reasonable people ought to coordinate on a particular political principle or rule, when they disagree about which political principles and rules are morally authoritative. Such a theory would merely justify the de facto authority of a political principle or rule that is *effective* in using coercion to achieve a stable political order. Without any appeal to reasonable people's moral reasons, or more precisely what they each see as moral reasons from their point of view, to obey a political principle or rule, a theory of political legitimacy would be unable to show how to achieve a stable political order amongst reasonable people. It would be unable to show why reasonable people ought not violently resist the authority. Rather, it would describe an oppressive order that reasonable people are justified in resisting.

To that end, any theory of political legitimacy that aims to take reasonable disagreement about justice seriously requires the broader view of political legitimacy I have outlined. It requires the view that political legitimacy is about both justifying political authority and the use of coercive power. In short, it involves reasonable people having sufficient moral reason to coordinate on a political principle or rule that is coercively enforced.

2.2 Disagreement to Legitimacy arguments

Having settled what political legitimacy is, we are now in a position to see what it means to theorise about political legitimacy. That is, what it means to argue for a theory of political legitimacy that takes reasonable disagreement about justice seriously

and help us achieve a stable political order. I have said that it involves a special sort of argument: a Disagreement to Legitimacy argument. Disagreement to Legitimacy arguments, as the name would suggest, are arguments that take disagreement as their premise and conclude with a theory of political legitimacy. To see what that means and what is distinctive about it, I propose the structure of a Disagreement to Legitimacy argument is as follows:

P1: Coordination on some political principle or rule that orders society is necessary for reasonable people to pursue their conceptions of the good life and resolve claim disputes.

P2: Reasonable people disagree about justice.

P3: Coercive coordination on a political principle or rule that ignores reasonable disagreement about justice is oppressive and gives people sufficient reason to violently resist.

P4: Some theory Y is the best explanation of reasonable disagreement about justice.

C : Given theory Y, the correct theory of political legitimacy is X.

Before unpacking the premises of the argument, I want to note that the argument is not strictly a deductive argument. It is an argument that reflects a process of reflective equilibrium where the conclusion is taking to be in equilibrium (ie. coherent) with the four premises. This means our evaluation of the conclusion should be a judgement about whether a particular theory of political legitimacy is actually in equilibrium with the array of considered judgements and social facts that make up the premises and not whether it is strictly derivable from the four premises.

The most important part of a Disagreement to Legitimacy arguments is its first three premises because they collectively establish a problem: *the problem of normative instability*.[4] The problem of normative instability is that when the facts in the first three premises hold, reasonable people do not have sufficient moral reason to coordinate, or maintain coordinating on a political principle or rule and they cannot achieve a stable political order when they desperately need to. In both cases, society

[4] I use the label "problem of normative instability" to group together similar and related problems that others like Gerald Gaus (2011b: 22–23, 261, 2011a: 306), Jonathan Quong (2011: 40), and Matt Sleat (2013: 74) have referred to as the, "two puzzles of moral authority", "problem of justificatory instability", "The Justificatory Problem", "the puzzle of how legitimate political principles are possible in light of the fact of reasonable pluralism", and the "problem of not taking pluralism seriously enough" respectively.

will be disordered because either there is no way to resolve claim disputes between reasonable people and their conflicting conceptions of the good life, or a political principle or rule is insufficiently justified so they have reason to violently resist.

The fourth premise and the conclusion of Disagreement to Legitimacy arguments collectively propose a solution to the problem of normative instability. The solution begins by working out why reasonable disagreement about justice – the source of normative instability – occurs and then proposing why in virtue of that, a particular theory of political legitimacy can solve the problem of normative instability. The solution then involves two parts. The first is the explanation of reasonable disagreement about justice because it describes what causes the problem of normative instability. Given reasonable disagreement *about justice* is merely a species of reasonable disagreement generally, and no theorist typically seeks to explain reasonable disagreement *about justice* in particular, the theories cited in this first part will typically be of reasonable disagreements in general or at least of any moral and political topic. These will then serve as the best explanations of reasonable disagreement about justice.

The second part of the solution is the theory of political legitimacy which is proposed as the correct normative response to the forces causing the problem of normative instability. On this view the theory of political legitimacy is best thought of as showing how it is possible, but not of course guaranteed, for reasonable people to achieve a stable political order. It is a point in general and wide reflective equilibrium coherent with the considered judgements implicit in the four premises.

But, why should we conceive of theorising about political legitimacy in this way? I propose two reasons. The first and most important reason is that Disagreement to Legitimacy arguments provide a clear framework for evaluating theories of political legitimacy fairly. In short, we evaluate them based on how well they solve the problem of normative instability within two fixed points. This means we evaluate the ability of a theory of political legitimacy to show how reasonable people's balance of reasons can provide and continue to provide sufficient moral reason to coordinate on a political principle or rule. We can make this more precise by evaluating how a theory fares on two dimensions: creating a political order and maintaining a political order.

Creating a political order is the ability of a theory to show how reasonable people can order their social world by coordinating on a particular political principle or rule in the face of reasonable disagreement about justice. This amounts to showing how there can be a balance of reasons amongst reasonable people such that they all have sufficient moral reason to coordinate on a political principle or rule.

Maintaining a political order is the ability of a theory to show how reasonable

people can maintain the political order in the face of both continued reasonable disagreement about justice, and the reasonable set of circumstances a society is likely to face over time. This amounts to showing how reasonable people's balance of reasons can be maintained in such a way that they continue to have sufficient moral reason to coordinate on a political principle or rule despite both exogenous and endogenous forces acting on their balance of reasons.

The framework also provides two fixed points for our reflective judgements when evaluating theories of political legitimacy across both dimensions. Theories that legitimate multiple conflicting political principles or rules do not solve the problem of normative instability because they do not create a political order for the whole of society. They result in social disunity. Also, theories that only make it possible for a society to successfully create and maintain a balance of reasons through coercion alone do not solve the problem of normative instability either. These would be theories where political principles or rules are legitimated because people fear coercion and not because they have sufficient moral reason to coordinate. Their reasonable disagreements about justice are quashed by the oppressive use of coercion. This means that the theory creates order in such away that it is disposed to fall into disorder because people always have a reason to violently resist their oppression.

These dimensions of evaluation and fixed points provide a suitably neutral way of describing the way theories of political legitimacy can solve the problem of normative instability even if they intend to solve it in different ways. The balance of reasons to coordinate can be cashed out as a strategic equilibrium, or as the outcome of non-strategic reflection on moral reasons. We need not at this stage be committed to a particular view of how people reason, their deliberative operations or how sufficient justification is achieved for every single reasonable person. All this will be part of any particular theory of political legitimacy that will be evaluated as we go.

The second reason why we ought to conceive of theorising about political legitimacy as making Disagreement to Legitimacy arguments is that it clarifies a contemporary debate between political liberalism and political realism as a first-order normative debate about political legitimacy. Since both views claim to take reasonable disagreement about justice seriously, conceiving of theorising about political legitimacy as making Disagreement to Legitimacy arguments shows how political liberals and political realists are actually in dialogue about a first-order normative issue. They are, on the view I have proposed, making competing Disagreement to Legitimacy arguments. They are in a first-order normative debate about theories of political legitimacy. A side effect of this will be that it guarantees political liberalism and political realism as

opposed to other theories about political legitimacy feature heavily in this thesis.

The existing debate between political realists and political liberals is a fractured one that hinges on vague generalisations and slogans about the relationship between morality and politics. For instance, political liberals claim to show how a political principle or rule can be justified as moral principles or moral rules and "stable for the right reasons" (Rawls 2005; Gaus 2011b; Quong 2011). But, political realists say that this is a form of "political moralism" that makes "the moral prior to the political". They claim to offer an alternative approach that "gives a greater autonomy to distinctively political thought" (Williams 2005; Galston 2010; Geuss 2008; Rossi and Sleat 2014). These slogans are then interpreted in different ways.

On one interpretation, theorists see the debate about political legitimacy to actually be a methodological debate about how to do political theory. Political realists claim political theory ought to be done with a distinct methodology, specifically, in a "practice-dependent" or "non-ideal" way (Sangiovanni 2008; Mason 2010; Finlayson 2017; Valentini 2012; Philip 2010, 2012). Some political liberals have responded that ideal theory can do the same job as the purported non-ideal realist theories, and that at worst ideal and non-ideal approaches to justice and legitimacy are complimentary (Estlund 2017; Erman and Moller 2013b, 2015a, 2016).

On another interpretation, theorists see the debate about political legitimacy to actually be a metanormative debate about the normativity involved in political theorising. Political realists claim that we ought to think of political theory as metanormatively autonomous from interpersonal morality. This means that political theory has its own distinct normativity and the normative force of political principles depends on facts and values that are completely different to that which moral principles depend on (Newey 2010; Rossi 2012; Galston 2010; Jubb 2019; Jubb and Rossi 2015a,b). Political liberals have responded in turn that this new distinct normativity is either functionally too similar to moral normativity in which case it isn't clear why moral normativity can't play the same role, is so distinct from moral normativity it is indistinguishable from instrumental normativity, or is grounded in controversial conceptual distinctions that are not shared by all political realists (Maynard and Worsnip 2018; Erman and Moller 2013a, 2015b, 2018).

I have no quarrel with interpreting the debate about political legitimacy between political liberals and political realism in these ways or for theorists who think those interpretations are important to carry on debating. But, I do not think these are the *only* ways to interpret the debate or that the debate about political legitimacy is limited to a methodological or metanormative debate. Disagreement to Legitimacy arguments

allow us to interpret the debate about political legitimacy between political liberals and political realists in a different way: as a first-order normative debate about the sufficient conditions for political legitimacy.[5] This is because Disagreement to Legitimacy arguments show how political liberals and political realists can have a common methodology and a common, or neutral, metanormative view.

Disagreement to Legitimacy arguments show that the way to theorise about political legitimacy can involve a methodology that both political liberals and political realists share. Specifically, first finding the best explanation of reasonable disagreement about justice and then proposing a theory of political legitimacy that, in light of that explanation, shows how reasonable people can achieve a stable political order within the two fixed points I mentioned previously.

Disagreement to Legitimacy arguments also show that the way to theorise about political legitimacy can involve metanormative commitments that political liberals and political can share. This is because Disagreement to Legitimacy arguments involve two assumptions. The first assumptions is that the normativity involved in theorising about political legitimacy is ordinary moral normativity or the type of normativity that reasonable people believe makes it the case that they ought to coordinate on coercively enforced political principles or rules. The second assumption is that there is no restriction on how reasonable people achieve conclusive justification to coordinate. This means that all the latitude that political realists want in the considerations that can speak in favour or against coordination are possible. Considerations can be potentially, agent-relative, agent-neutral, context-relative, context-neutral, or some mix of all four. Disagreement to Legitimacy arguments allow us to see political liberals and political realists as engaging in a first-order debate where there is enough normative tools for each to propose a solution to the problem of normative instability.

All this means that, by conceiving of theorising about political legitimacy as making Disagreement to Legitimacy arguments, we can interpret the debate between political realists and political liberals as a first-order normative debate. They are both making Disagreement to Legitimacy arguments, but propose different theories of political legitimacy given the premises. I believe this clarifies how many political liberals and political realists theorise about political legitimacy.

To see this I submit we can reconstruct what some prominent political liberals and political realists say when arguing for their theories of political legitimacy in the

[5] See Rossi (2019) on "ordorealism" and Jubb (2015) for political realists who interpret the debate as I do.

form of a Disagreement to Legitimacy argument:

> Premises 1-3: For political liberals, the three premises are explicated in three distinct ideas: the "objective circumstances of moderate scarcity", the "fact of reasonable pluralism", and the "fact of oppression" (Rawls 2005: 36–37, 66).[6] For political realists, the three premises are explicated collectively in two ideas: the "circumstances of politics" which includes the "radical and permanent political disagreement" that makes coordination necessary, and the idea that politics cannot merely involve "successful domination" or in other words oppressive coercion (2013: 15, 46–57, 113–114; 2005: 3, 77).[7]
>
> Premise 4: Political liberals cite the "burdens of judgement" as the best explanation of reasonable disagreement (Rawls 2005: 54–58).[8] Political realists tend to either agree with political liberals or cite the "obscure mixture of beliefs (many incompatible with one another), passions, interests" as the best explanation (2013: 133–134; 2005: 13).[9]
>
> Conclusion: Political liberals conclude that a theory with "public justification" as its core normative standard is the correct theory of political legitimacy (Rawls 2005: 70, 387–388).[10] Political realists conclude that a theory with "meeting the Basic Legitimation Demand" or an "acceptable solution to the first political question" as its core normative standard is the correct theory of political legitimacy (2013: 152–153; 2005: 4).

Seen this way, the debate between political liberals and political realists is at a first-order normative level about the legitimacy of political principles or rules that are coercively enforced.

Given all that, a Disagreement to Legitimacy argument is a plausible way to interpret what a political liberal like Rawls (2005: xix) is saying when he writes:

> Given the fact of the reasonable pluralism of democratic culture, the aim of political liberalism is to uncover the conditions of the possibil-

[6] See also Gaus (2016: Ch. 3, 2011b: Ch. 1, 1999), Quong (2011: 36).

[7] See also Waldron (1999: 101–103).

[8] See Quong (2011: 37).

[9] See also McQueen (2018: 256–260) and Hall (2017: 285–286).

[10] See also Gaus (2016: Ch. 4, 2011b: Ch. 5, 1999), Quong (2011: 180–191).

> ity of a reasonable public basis of justification on fundamental political questions.
>
> It should, if possible, set forth the content of such a basis and why it is acceptable. In doing this, it has to distinguish the public point of view from the many nonpublic (not private) points of view. Or, alternatively, it has to characterize the distinction between public reasonable and the many nonpublic reasons and to explain why public reason takes the form it does (VI).

In the same vein, a Disagreement to Legitimacy argument is a plausible way to interpret what a political realist like Sleat (2013: 47) is saying when he writes:

> The persistence of disagreement is one of the fundamental and 'stubborn facts' of political life which ensures that there is rarely any natural harmony or order in human affairs. The most basic political question, what I shall call 'the political question', is how we are to live together in the face of such deep and persistent disagreement. The primary objective of politics must therefore be to provide a framework that creates order and stability by establishing the terms on which we are to co-exist and also the means for making future commonly binding decisions in conditions of disagreement (including the procedures for altering the terms of co-existence). Any successful answer to the political question will therefore require a structure of institutions and practices that provides the basis for persons to live together under a common political authority.

In sum, conceiving of theorising about political legitimacy as making Disagreement to Legitimacy arguments shows that political liberals and political realists are engaged in the same project. They are both offering, in light of an explanation of reasonable disagreement about justice, competing theories of political legitimacy aimed at solving the problem of normative instability.

The upshot of all this is that it shows why much of this thesis will focus on evaluating various political liberal and political realist conceptions. Given the goal of this thesis is to argue for a particular political realist conception of political legitimacy – the Dual Convergent Conception – much of it, specifically Chapters 4 and 5, will be devoted to showing how extant conceptions of political liberalism and political realism do not solve the problem of normative instability. This will provide the negative argument in favour of the Dual Convergent Conception.

2.3 Reasonable Disagreement about the Good Objection

One objection to my characterisation of theorising about political legitimacy as making Disagreement to Legitimacy arguments might be, that the interpretation of normative instability is wrong. This is because it is not consistent with what political liberals, who I claim make Disagreement to Legitimacy arguments, are trying to solve. The idea is that the most prominent version of political liberalism – a Rawlsian conception – is premised on how disagreement about *the good* makes achieving a stable political order a problem and not disagreement about *justice*. As Paul Weithman (2015: 75, 83–88, 2010: 319–321) and Jonathan Quong (2011: 36–37, 137–138) argue, this is the view that the problem political liberalism seeks to solve is internal to liberal political theory. Liberal institutions permit diverse conceptions of the good and so produce reasonable disagreement about the good. The question is then how can liberal political theory justify itself to people whor are committed to liberal values and yet disagree about what is valuable in human life. How can it garner and maintain their support once they go about their lives pursuing what they view as the good? This is the stability problem that Rawls unsuccessfully tried to solve in Part III of *A Theory of Justice* and that *Political Liberalism* was aimed at solving. None of this involves denying reasonable disagreement about justice exists, but rather that the central animus for the political liberal's theory of political legitimacy is disagreement about the good. As such a Disagreement to Legitimacy argument is the wrong way to conceive of a political liberal's argument for their theory of political legitimacy.

I believe this objection fails in at least two ways. The first way the objection fails is that it rests on a contested interpretation of the Rawlsian conception of political liberalism. The second way the objection fails is that it leads to a dilemma between requiring an implausible reading of the "fact of reasonable pluralism" in Rawls's conception of political liberalism and relying on a plainly false view of what makes a theory of political legitimacy necessary.

To the first way it fails, other political liberals like Gerald Gaus (2017: 27–30, 2014: 243–248), Kevin Vallier (2019: 5–7), David Thrasher (2018: 399–400), Brian Kogelmann (2017), David Reidy (2007: 250–251), and Burton Dreben (2003: 320–321) do not understand Rawls's conception as arguing for a theory of political legitimacy purely on the basis of reasonable disagreement about the good. Rather they see it as argued for on the basis of reasonable disagreement about justice which will include disagreements about the right and the good to varying degrees. To that extent, I do not think the interpretation I have offered breaks significantly with the broader tradition of interpreting the Rawlsian conception of political liberalism.

However, independent of the interpretation of other political liberals, the objection fails for another more worrying reason. It faces the following dilemma. On one horn the objection relies on a reading of Rawls's notion of the "fact of reasonable pluralism" as disagreements involving *only* conceptions of the good. This is what Rawls would have to have thought about reasonable disagreement for the objection to go through. If this were the case, then Disagreement to Legitimacy argument would indeed be the wrong way to think about Rawls's project. But, this is an implausibly narrow reading of reasonable pluralism for at least two reasons.

First, as Rawls (2005: xvii, 36–37) says the "fact of reasonable pluralism" involves conflicting *comprehensive religious, philosophical, and moral doctrines*. These are doctrines that are both general and comprehensive. Which means they apply to a wide range of subjects and involve a wide array of values. Rawls specifically contrasts them with *political conceptions of justice* which are neither general nor comprehensive. Rawls does not contrast them with political conceptions of the good, or some other non-comprehensive doctrine of the good. The idea is that comprehensive doctrines lead us to derive conceptions of right and conceptions of the good that tell us how we ought to conduct our lives and what is ultimately of value in all domains of life. As such they yield comprehensive conceptions of justice that order our social life in total according to a wide array of values and with regard to all domains of life.[11] Therefore, simply as a matter of what Rawls contrasts the content of reasonable pluralism with, we have good reason to think it involves disagreement about justice and not merely disagreement about the good.

The second reason why the objection would rely on an implausibly narrow reading of the fact of reasonable pluralism is because of the examples that Rawls uses to illustrate comprehensive doctrines. When considering the comprehensive doctrines that lead to the problem of normative instability, Rawls (2005: xviii, xx, 346, 489–490) includes religious conceptions of justice like Catholic and Protestant Christianity, and nonreligious conceptions of justice like utilitarianism, Millian and Kantian liberalism, and socialism. These are doctrines that conflict on more than the good, they also conflict about the right. As such, they would be involved in disagreements about justice – about how to order a society according to a scheme of rights, opportu-

[11] See Gaus (2004) for helpful clarification of how comprehensive doctrines are related to conceptions of justice. Further evidence for this is that for Rawls (2005: 173–174) the good and the right are complementary. This means a comprehensive doctrine of justice will contain both normative and evaluative content intertwined with each other. As such, a comprehensive doctrine will at the very least have to include a thin notion of justice that pertains to what it is right to do even if it is merely 'you ought to always maximise the good or pursue what is necessary for realising the good'.

nities, and resources amongst individuals, institutions and social systems – that might stem from underlying normative or evaluative differences.

If the objectors abandon the reading of "fact of reasonable pluralism" as disagreements involving only conceptions of the good, then they face the second horn of the dilemma. Which is that if they want to maintain the objection they are committed to saying that the only disagreements that matter for theorising about political legitimacy are disagreements about the ideas of the good in a conception of justice. The idea being that only disagreements about the good make a theory of political legitimacy necessary. But, this is plainly false. Other aspects of a conception of justice, namely ideas of the right, can also make a theory of political legitimacy necessary. Disagreement about the principles or rules it is right to coordinate on and coercively enforce, combined with agreement about the good to be furthered, would be enough to make a theory of political legitimacy necessary. The use of coercive power would be, as Rawls (2005: 37) is prescient to point out, oppressive in trying to create a society unified under a single comprehensive doctrine. Insofar as conceptions of justice include ideas of the right and the good, disagreement about the ideas of the right are as capable of making political legitimacy necessary as much as disagreements about the good.

To that end, Disagreement to Legitimacy arguments actually provide a better picture of the sort of disagreement that matters for theorising about political legitimacy. It shows how disagreement about justice, which can include disagreement about ideas of the right and the good, can play a role in making a theory of political legitimacy necessary. In sum, the objection fails against my construal of theorising about political legitimacy through the idea of Disagreement to Legitimacy arguments.

2.4 Fact-Sensitivity Objection

Another objection one might have to my characterisation of Disagreement to Legitimacy arguments is that whatever they conclude with they cannot be held to argue for fundamental normative principles about political legitimacy. This is because, taking inspiration from Gerald Cohen's (2008: 236–244) argument against Rawlsian principles of justice, one might point out that fundamental normative principles are fact-insensitive. But, the whole point of Disagreement to Legitimacy arguments involves arguing for a theory of political legitimacy on the basis of how it responds to the best explanation of reasonable disagreement about justice. In Cohen's language this would be a theory that is sensitive to the facts that explain why reasonable dis-

agreement about justice occurs.[12] But, the explanation of why these facts matter will involve a further principle, a fundamental normative one. This principle would be something like: 'a society ought to care about the best explanation of reasonable disagreement about justice to avoid the consequences of normative instability'. This is the fundamental normative principle of political legitimacy that must be argued for. This means that what Disagreement to Legitimacy arguments conclude with cannot be a theory of political legitimacy. They, at best, argue for regulative principles that *serve* the fundamental normative principles of political legitimacy.

I have no real response to this objection other than to make two remarks. One about philosophical methodology and the second about the importance of making Disagreement to Legitimacy arguments. The first remark is that the Fact-Sensitivity Objection seems to rest on a fundamental difference in philosophical methodology. As Cohen (2008: 3) concedes directly, "...there is a disagreement about how to do political philosophy, or indeed philosophy itself". Much of this difference boils down to what I take the aim of political theory to be, namely to formulate normative principles that tell us how to live with others in a shared social world. Given that, political theory is free to take into account facts about our shared social world to help us theorise about normative principles.[13] As long as this theorising is done correctly and the facts we take into account like the best explanation of reasonable disagreement about justice are actually relevant for the problem at hand, then the principles we arrive at will ipso facto be fundamental normative principles. To that end, I believe fundamental normative principles can be fact-sensitive and there is nothing wrong with this.

The second remark, since my first remark is unlikely to satisfy objectors, is that I do not think the Fact-Sensitivity Objection provides any good reason to stop making Disagreement to Legitimacy arguments even on the objectors grounds. This is because even if I was to concede that Disagreement to Legitimacy arguments ultimately only conclude with regulative principles that serve fundamental principles of legitimacy, it is still worthwhile making them. This is because even if we accept the view that the fundamental normative principle at work in Disagreement to Legitimacy arguments is the principle that "a society ought to avoid the consequences of normative

[12] This is despite Cohen's (I believe mistaken) view that the Rawlsian consensus principle of legitimacy is *fact-insensitive* (Cohen 2008: 297–298). Cohen seems to ignore that Rawls proposes his principles precisely in light of his explanation of reasonable pluralism that his pure constructivist method for arriving at principles of justice could not.

[13] See Larmore (2013) and Miller (2013: Ch. 2) on this way of doing political philosophy. I also defend this view of political theory elsewhere ("The Independence of Political Theory": ms).

instability", this does not help us choose normative principles that will actually help us live with others in a shared social world. That normative principle is fairly uncontroversial and is shared by many political theories. The interesting and more pressing question is how does a society actually go about avoiding the consequences of normative instability. For this we need to make Disagreement to Legitimacy arguments and find a theory of political legitimacy. If one wants to think about such a theory as containing non-fundamental normative principles then so be it. But, this does not detract from the importance of trying to theorise about them.

3 Reasonable Disagreement about Justice as an Explanandum

In the previous section I described the general methodology of this thesis. I described what I will take political legitimacy to be and that Disagreement to Legitimacy arguments is how we ought to conceive of theorising about political legitimacy which aims to take reasonable disagreement about justice seriously. I also said that one upshot of adopting this methodology is that it means that the first step of arguing for a theory of political legitimacy is engaging in an explanatory project to find the best explanation of reasonable disagreement about justice. This was because such an explanation will tell us what causes the problem of normative instability and so allow us to formulate a theory of political legitimacy that responds correctly to this cause. To that end, in this section I clear the ground for engaging in such an explanatory project. I will establish what the explanandum is for such an explanatory project. In short, I will argue for a specific view of what reasonable disagreement about justice is.

All this might seem confusing. At the outset I said that the point of this thesis would be to argue for a novel political realist conception of political legitimacy, namely the Dual Convergent Conception. But, one of the reasons I gave in favour of Disagreement to Legitimacy arguments was that they clarify a debate that has emerged in recent years between political liberals and political realists. In doing so I showed that political realists already offer what they take to be the best explanation of reasonable disagreement about justice (ie. they either agree with political liberals about the burdens of judgement or refer to reasonable people's passions and emotions). What more is there to do?

Well, as I'll argue in Chapter 2, there are good reasons to think extant explanations of reasonable disagreement about justice, like the burdens of judgement, are defective. As I'll show, current explanations offered by political liberals and political realists are

inadequate. They are either poorly argued for, do not separate political liberals and political realists, or are too vague and general to apply to real cases. In fact, this will form one part of a broader argument against all current ways of explaining reasonable disagreement about justice. This will culminate, in Chapter 3, with an argument for a novel explanation. But, before I do any of this we need to be clear on what is being explained, namely what is the explanandum of such an explanatory project.

In formulating reasonable disagreement about justice as an explanandum I start with the thought that it must at least be prima facie fair in two ways. The first way the explanandum should be fair is that it does not so heavily influence the explanatory project to come that it ensures, before any evaluation, a particular explanation. This can happen because a formulation of reasonable disagreement about justice can be so precise to contain its own story of why it occurs. For instance, reasonable disagreement about justice could be defined in such a way that it is whatever disagreement that is the result of X sort of judgement making. This would undermine the point of the explanatory project from the start. The second way the explanandum should be fair is that it should not be designed to provide a rationale for any particular theory of political legitimacy. This would undermine the explanatory project altogether and the evaluation of Disagreement to Legitimacy arguments. If, as I have argued we ought to conceive of theorising about political legitimacy as Disagreement to Legitimacy arguments, then we want the normative appeal of a theory of political legitimacy to depend on how much better it is at solving the problem of normative instability than its competitors. We do not want it be guaranteed beforehand. To that end, I propose the way we ought to formulate reasonable disagreement about justice is to have an explanandum that, at most, makes a theory of political legitimacy necessary. In light of that, I submit that to explain reasonable disagreement about justice is to explain:

Reasonable Disagreement about Justice: A state of affairs of intractably conflicting judgements about the institutions and outcomes justice requires, made by at least two parties who both have, a minimal capacity for rationality and a minimal capacity for sincerely making judgements that they think others can agree to.

This formulation of reasonable disagreement about justice makes a theory of political legitimacy necessary because it involves five important features. The first three concern the type of conflict involved in the disagreement, and the last two concern the type of people involved in the disagreement.

The first feature is that, the disagreement has to involve a conflict in the content of people's judgements. This means the disagreement involves a state of affairs of two

people making judgements that are incompatible, such that they both cannot be true, with respect to a particular object of thought.[14] This ensures that the disagreement meets a minimum threshold of genuineness and is not merely an illusory disagreement where reasonable people's judgements are actually compatible. Without this feature, the entire premise of this thesis would not get off the ground. Rather than looking for a theory of political legitimacy we would need to sort out how to label things or work out why a disagreement is illusory.

The second feature is that, the disagreement has to involve conflicting judgements about a particular subject: the institutions and outcomes that justice requires. Before seeing why let me clarify what I mean by this. I mean that the judgements involved in the disagreement have to be those where the object of the judgements are institutions and outcomes, and the criterion of judgement is justice. By "institutions" I mean the constitutions, ordinary statute laws, and policies which Rawls (2005: 11–12) referred to as the basic structure of a society. By "outcomes" I mean the distribution of rights, opportunities and resources that those institutions produce. These institutions and outcomes are then evaluated according to people's beliefs about what justice requires, which results in a judgement about the institutions and outcomes that justice requires. It is important to get clear about this because what I mean by "conflicting judgements about the institutions and outcomes that justice requires" is that people conflict in their evaluation of institutions and outcomes in virtue of their conflicting beliefs about what justice requires.[15] This means they make conflicting judgements not in virtue of conflicting evaluations about what institutions and outcomes satisfy an agreed upon set of beliefs about the requirements of justice. Rather they make conflicting judgements in virtue of conflicting beliefs about the requirements of justice are.

Reasonable disagreement about justice has to involve the sort of conflicting judgements I have described above. Otherwise, there would be no need for a theory of political legitimacy. If people merely disagreed empirically about the policies or laws that best satisfy what justice requires, they would have two options open to them. On the one hand they could simply appeal to the procedure that justice requires – which recall by hypothesis they both agree on – for resolving disagreements about

[14] See Frances (2014: Ch. 1) and Gibbard (2003: Ch. 4) for this basic way of thinking about genuine disagreement.

[15] See Valentini (2013: 183–187) for a helpful distinction between "thick disagreement about justice" and "thin disagreement about justice" which corresponds to the distinction I make here. Although Valentini ends up with a slightly different and more general formulation of reasonable disagreement, the core distinction matters both for her account and mine.

the policies and laws. On the other hand they could simply spend more time and effort explaining their evaluation of how some policies or laws satisfy the requirements of justice to reach agreement. In both cases, they would then agree about the principles or rules that people ought to obey and that can be coercively enforced.[16] A theory of political legitimacy and its claims about when reasonable people have sufficient moral reason to coordinate on a political principle or rule that is coercively enforced only becomes necessary when the options I mentioned are unavailable. It is only when parties disagree about the institutions and outcomes that justice requires that they require a theory which explains how reasonable people can have sufficient moral reason to coordinate on a political principle or rule that is coercively enforced.

The third feature is that, the conflict in judgements over the institutions and outcomes that justice requires have to be *intractable*. This means that the disagreement persists despite the disagreeing parties explaining their beliefs about what justice requires and their reasons for their judgements. Otherwise, we would not need a theory of political legitimacy but rather more time for reasonable people to have their disagreement and convince each other that their judgement about which institutions and outcomes justice requires is correct. It is only when people need to realise some institutions and distributive outcomes to order society but cannot because their disagreement about justice persists despite their best efforts to resolve it, that a theory of political legitimacy becomes necessary. Reasonable people require a theory to tell them how their disagreement can be resolved or accommodated in such a way that they can achieve a stable political order.

The fourth feature is that, the disagreement has to involve parties who both possess a minimal capacity for rationality because otherwise the person with the lower capacity for making rational judgements should defer to the one whose capacity does meet the minimal threshold for rationality.[17] This is because if we assume, according to the definition, that they both have a minimal capacity to sincerely make judgements that they think others can agree to, then the person with the lower capacity for rationality ought to defer. This is because part of being sincere is to see that the person that meets the minimal capacity threshold for rationality is more disposed to getting

[16] But, note this does not exclude the fact parties might be having a reasonable disagreement about the requirements of justice *because* of some foundational empirical disagreement.

[17] See, for similar construals of this aspect of reasonable disagreement, Gaus (2011b: 276–277) on a "basic level of reasoning" which also allows for the possibility for reasonable people to engage in more or less sophisticated reasoning above this level, and McMahon (2016: 61–66) on "reasonableness in the competence sense" which involves the idea that "A person is reasoning competently in a particular case when his drawing of a conclusion, or generating a cognitive product of some other kind, manifests the proper functioning of the relevant mental capacities".

the right answer. It would be insincere of a person to insist on their judgement of a matter when they know that they do not even meet the minimal threshold capacity for rationality. This is because the person who does meet the threshold is more likely to be sensitive to the moral reasons in the context and make a coherent calculation of them. What ought to happen is that in deferring the person who meets the minimal threshold capacity should explain how they deliberate, came to their judgement and why the other has made a mistake.[18]

The fifth feature is that, the disagreement has to involve parties who both have a minimal capacity for sincerely making judgements that they think others can agree to, otherwise the entire point of taking reasonable disagreement about justice seriously is undermined.[19] This is because reasonable disagreements would include disagreement between at least two kinds of people. One kind of person is the contrarian who disagrees for the sake of disagreeing and not because they sincerely believe in the normative force of what they have judged justice requires. Since there is no amount of agreement or disagreement that will move them to coordinate, either we are justified in coercing them to avoid the consequences of not coordinating or we are justified in concluding they do not really possess the minimal capacity for rationality because they do not see the devastating consequences for social life if they do not coordinate on a political principle or rule.

The other kind of person that would be included is the fundamentalist who insists on the political principle or rule they believe justice requires no matter the circumstances. This person is subtly different from the contrarian because whilst they do sincerely believe what they say, they do not make judgements they believe others can agree to. They make judgements that do not appropriately respond to the fact that they must try and live with others in a shared social world. Perhaps they do not try and explain the reasons for their judgement or even try to bring themselves to see their interlocutor's point of view. They are making judgements about justice in bad faith to dominate or gain power over their fellow citizens. In these cases, it seems coercion is pointless because a stable political order is nigh impossible to create. Rather

[18]Note even in this situation one party could object that justice requires that people with greater capacity ought not coerce those with a lower capacity. This would merely show that the parties now suddenly do face a reasonable disagreement about justice for which they require a theory of political legitimacy.

[19]See Gaus (2011b: 276, 288–292) and Carey (2018: 51–59) for a similar minimalist sincerity condition as opposed to the far more demanding notion proposed by Quong (2011: 265–273), and McMahon (2016: 66–73) on "reasonableness in the concession sense" which involves a similar idea of making judgements with a disposition "to respond to perceived disparities of concession by making or seeking corrective concessions, provided that others are similarly disposed".

the parties should go their separate ways. A theory of political legitimacy has nothing to recommend in such cases.

In sum, reasonable disagreement about justice is disagreement between those of a certain moral character. It is disagreement between those who meet a minimal threshold in capacity for reasoning and thinking about justice, and who in good faith want to live in a shared social world with others. In short, it is disagreement between morally decent people such that neither can claim to coerce the other without some further justification.

It is worth clarifying at this stage how the explanandum of reasonable disagreement about justice I have formulated differs from other notions of disagreement used in philosophy and why those other notions would not do for our purposes. Firstly, reasonable disagreement about justice is not faultless disagreement.[20] This is disagreement that is assumed to be about a topic that is beyond rational conflict. As such it contains its own explanation, namely that it is caused by non-truth apt judgement making. Although this is a conclusion we might arrive at by engaging in the explanatory project, it would not be fair to assume it beforehand.

Secondly, reasonable disagreement is also not peer disagreement.[21] This is disagreement between parties that are actually epistemically equal in all regards. This would be an implausible picture of disagreements about justice. Society is obviously made up of people with varying epistemic abilities and so idealising reasonable disagreement to those that are equal in their epistemic abilities and have all the same evidence would at worst involve no actual disagreement at all and at best involve an implausibly small number of actual disagreements to be worth explaining.[22]

Finally, reasonable disagreement about justice is not disagreement between people committed to intellectual modesty.[23]. This is the sort of disagreement some theorists have tried to formulate where the very property of reasonableness is designed to provide a justification for political liberalism. This would be unfair because it undermines the entire point of trying to find the best explanation of reasonable disagreement about justice and evaluating competing theories of political legitimacy in light of that explanation. After all we would already have a perfectly good reason to prefer one theory of political legitimacy over another.

[20] See Kolbel (2004) on the details of this type of disagreement.

[21] See Kelly (2005) and Gutting (1982) on the details of peer disagreement and, Peter (2013) and van Wietmarschen (2018) on its use in political liberalism.

[22] See Frances (2014: 166), King (2012), and Matheson (2014: 320–328) on similar points about the irrelevance of peer disagreement.

[23] See Leland and van Wietmarschen (2012) for this sort of way of formulating reasonable disagreement with regard to political liberalism.

All this achieves the goal of this section which was to lay the groundwork for the explanatory project in this thesis in a fair way. Reasonable disagreement about justice as I have described it is prima facie neutral and fair between different explanations of it and different theories of political legitimacy which will be tested by the best explanation of it. It is in simple terms a neutral way of describing what Rawls (2005: 55) describes as reasonable disagreement: disagreement between people who have a capacity for a sense of justice and a conception of the good to an sufficiently equal degree. It does not presuppose any controversial feature of people or disagreement that advantages any particular explanation or particular theory of legitimacy in some pre-theoretic way. To that end, it is the suitable explanandum for the explanatory project in this thesis. It describes what impedes stable political order and so makes a theory of political legitimacy necessary.

4 Road Ahead

In this chapter I argued for a particular view of what it means to argue for a theory of political legitimacy that takes reasonable disagreement about justice seriously. I first settled the details of what I take political legitimacy to be on the basis that only a broad view of political legitimacy that includes both the justification of political authority and the use of coercive power can show us how to achieve a stable political order. I then argued that theorising about political legitimacy involves making Disagreement to Legitimacy arguments on the basis of two reasons. The first reason is that Disagreement to Legitimacy arguments provide a clear framework for evaluating theories of political legitimacy fairly. The second reason is that the arguments clarify a contemporary debate between two views of political legitimacy both of which claim to take reasonable disagreement about justice seriously.

An upshot of all that was that it showed that arguing for a particular theory of political legitimacy involves two stages. First, finding the best explanation of reasonable disagreement about justice, and second, arguing that in light of that explanation a particular theory of political legitimacy does best at solving the problem of normative instability. The rest of this thesis takes up these two stages. In Chapter 2, using the explanandum I argued for in §3, I evaluate competing explanations. I conclude that all extant explanations are found to be lacking. In Chapter 3, I propose a novel explanation – the Diverse Packages Theory – and argue that it does better as an explanation than extant explanations.

With the Diverse Packages Theory as the best explanation of reasonable disagree-

ment about justice in hand, I then begin the second stage of the thesis. This involves a negative and positive argument for a novel theory of political legitimacy: the political realist Dual Convergent Conception. In Chapter 4, I argue that in light of Diverse Packages Theory, no conception of political liberalism is a satisfactory theory of political legitimacy because they fail on at least one of the two dimensions of normative stability. This I argue motivates a general shift to political realism. In Chapter 5, I argue that in light of the Diverse Packages Theory, no extant conception of political realism is a satisfactory theory of political legitimacy either. In Chapter 6, I propose the Dual Convergent Conception avoids the problems of its competitors and manages to show how reasonable people can create a stable political order. That it, unlike its competitors, takes reasonable disagreement about justice seriously and as a result shows how reasonable people ought to act from the perspective of eternity.

Chapter 2

Explaining Reasonable Disagreement about Justice

1 Introduction

This chapter begins the first part of my Disagreement to Legitimacy argument. It begins the explanatory project of finding the best explanation of reasonable disagreement about justice. In the previous chapter I cleared the ground for this by settling the details of what it means to take reasonable disagreement about justice as an explanandum. In this chapter I evaluate extant theories that aim to explain that explanandum and argue in favour of a particular type of theory: Concept Possession and Use theories. I show that it both avoids the problems that beset the other extant theories and is the most explanatorily powerful. But, I conclude that the only extant theory that instantiates it has some serious flaws that should worry us. This will eventually lead to my argument in Chapter 3 where I propose an alternative instantiation of it: Diverse Packages Theory.

The chapter proceeds as follows. In §2, I detail two enrichments to the theoretical landscape of theories that will frame the argument in this chapter and the next. The first involves settling the idea of explanation that I will use for evaluating theories and the second involves settling the taxonomy of theories that will be considered.

In §3, I consider the Imperfection Family of theories that explain reasonable disagreement and argue that the two types of theories in this family face some serious problems. I argue that they either, rule out the existence of reasonable disagreement itself, cite facts that do not make a difference between disagreement and agreement, or fail to justify the normative standard that would pick out the fact that does make a difference.

In §4, I consider the Historical-Psychological Family of theories and argue that the only type of theory in this family that actually aims to explain reasonable disagreement also faces some series problems. I argue that it, either cannot explain or offers an implausible explanation of disagreements that arise because a person changes their judgement through reflecting on the deliberative process. I conclude that this is because these types of theories rely on a specific theory of moral psychology with insecure empirical grounds.

In §5, I consider the Conceptual Family of theories. I argue that one type of theory in this family – Concept Use theories – whilst avoiding the objections that faced the Imperfection and Historical-Psychological Family of theories faces a serious problem. It inherently cannot explain reasonable disagreements that are deep disagreements and so is explanatorily weak. I argue, however, that another type of theory – Concept Possession and Use theories – does better and therefore stands out as comparatively the best explanation of reasonable disagreement. This is because it avoids the objections that face the Imperfection and Historical-Psychological Families, *and* can explain reasonable disagreements that are deep disagreements. Given that, it offers the most powerful explanation of reasonable disagreement. But, I argue, this comes at the cost of two unique problems. I argue that, rather than giving us reason to reject Concept Possession and Use type of theories, it motivates us to look for a better instantiation of it, which I take up in the next chapter.

2 Two Enrichments

The argument in this chapter is framed by two enrichments of the theoretical landscape. In this section I lay out both of these enrichments. In §2.1 I clarify what it takes for a theory to count as giving an explanation of reasonable disagreement about justice. Specifically, I detail what it means to cite difference makers in order to explain reasonable disagreement about justice. This sets the terms on which different views will be evaluated. In §2.2 I lay out the taxonomy of theories I use in this chapter. Specifically, I detail how this is an expanded taxonomy from the one found in the extant literature and how it will involve three families of theories: the Imperfection Family, the Historical-Psychological Family and the Conceptual Family.

2.1 The Idea of Explanation

Explanations of any kind are judged by how well they explain the explananda they target. When sorting and evaluating theories that explain anything, what is relevant

is the different ways they explain a given explanandum. This is no different for explaining reasonable disagreement about justice. To that end, recall from Chapter 1 the explanandum at the centre of this thesis:

Reasonable Disagreement about Justice: A state of affairs of intractably conflicting judgements about the institutions and outcomes justice requires, made by at least two parties who both have, a minimal capacity for rationality and a minimal capacity for sincerely making judgements that they think others can agree to.

Now as I said in the previous chapter, reasonable disagreement *about justice* is merely a species of reasonable disagreement generally, and no theorist typically seeks to explain reasonable disagreement *about justice* in particular. This means that evaluating theories that explain the explanandum above will involve evaluating theories that aim to explain reasonable disagreement in general or at least of any moral and political topic. Such theories will then by definition explain reasonable disagreement about justice.

With that out of the way, the explanandum above, given it is a state of affairs, suggests a particular type of explanation, namely a causal explanation. In simple terms 'evaluating theories' is a matter of evaluating theories' competing causal explanations of reasonable disagreement. A state of affairs is explained by a causal process or sequence of states of affairs that lead up to it. This amounts to explaining the causal process that leads reasonable people to make the sort of intractably conflicting judgements that comprise reasonable disagreements about justice.

But, what does it actually mean to explain a state of affairs by describing the states of affairs that lead up to it? It means offering a contrastive explanation that describes what makes the difference between the state of affairs to be explained – the explanandum – being realised and a state of affairs that is not realised – the contrast class.[1] For reasonable disagreement the contrast class is the state of affairs where reasonable people make the same judgement. This is, in short, reasonable *agreement*.

Contrastive explanations cashed out in terms of describing 'difference makers' involves making counterfactual dependence claims. In the counterfactual dependence claims about reasonable disagreement the antecedent will be a negation of the proposed difference making state of affairs and the consequent will be a negation of reasonable disagreement or in other words what is in the contrast class for reasonable

[1] In describing what it means to explain a moral phenomenon like reasonable disagreement about justice, I rely on the general picture of causal explanations advanced by James Woodward (2003: 9–12) and Bas Van Fraassen (1980: 134–157). This does not commit me to an particular view of scientific explanation, but rather only to a general idea of causal explanations appropriate for the kind of task this thesis aims to do in moral philosophy.

disagreement. All this amounts to a contrastive explanation that describes the state of affairs that makes the difference between reasonable disagreement and reasonable agreement.

I submit that a contrastive explanation that describes differences makers is the idea of explanation that ought to concern us because it is the sort of explanation that allows theorists to make Disagreement to Legitimacy arguments. The point of the explanatory project, as I mentioned in Chapter 1, was to work what causes the problem of normative instability and therefore, allow us to evaluate which theories of political legitimacy do best at dealing with it. If an explanation of reasonable disagreement about justice is really going to help us in that evaluation then what is most relevant for us is that the explanation give us an account of why reasonable disagreement about justice is instantiated rather than not. And this is what a contrastive explanation that describes differences makers does. It captures the dependence between some particular states of affairs – the difference makers – and the instantiation of reasonable disagreement about justice.

2.2 The Expanded Taxonomy

The idea of explanation I have detailed suggests a taxonomy where theories belong to different families based on the sort of difference makers they cite for explaining why reasonable disagreement occurs. One starting point could be the taxonomy introduced by Andrew Mason (1993).[2] Mason (1993: 2–3) divides theories into two families: the "Imperfection conception" and the "Contestability conception". Using that taxonomy, the Imperfection conception would include theories that explain reasonable disagreement by saying reasonable disagreement occurs because at least one of the disagreeing parties makes their political judgement using "defective reasoning". In contrast, the Contestability conception would include theories that explain reasonable disagreement by saying that reasonable people use political terms which allow for different applications, without any form of intellectual error, as long as there is some freedom of expression (Mason 1993: 3). But, this division of families is too coursegrained to distinguish between theories that give different explanations for *why* "political terms allow for a variety of different applications".

To deal with this I suggest a taxonomy of three families: Imperfection, Historical-

[2] To be precise Mason evaluates explanations of 'political disagreement' in general. But, his account will do as a relevant starting point since explanations of reasonable disagreement are a subset of the explanations he is concerned with. Moreover many of Mason's examples are of how theories explain reasonable disagreements (Mason 1993: 7–12, 117–119).

Psychological and Conceptual. Whilst keeping Mason's original category of the Imperfection family of theories, I suggest dividing the Contestability family into the Historical-Psychological and Conceptual families. This means that theories that in the Imperfection family argue for facts about intellectual error as the difference makers. Theories in the Historical-Psychological family explain *why* "political terms allow for a variety of different applications" by citing particular token psychological and historical facts about reasonable people which affects their moral judgement making, as difference makers. Theories in the Conceptual family explain *why* "political terms allow for a variety of different applications" by citing facts about the role of concepts in the cognitive process of moral and political judgement making, as difference makers.

It is important to note that theories in the Historical-Psychological family should not be thought of as concerned with any and all historical and psychological facts that might underwrite an individual's moral judgement making. After all there is a sense in which all explanations involve historical and psychological facts. If reasonable disagreement about justice is a state of affairs of conflicting judgements about the institutions and outcomes justice requires, then it involves mental states causing other mental states over time. Any explanation of why these judgements conflict will involve reference to historical and psychological facts. This will be true for theories in the Imperfection and Conceptual families. Facts about intellectual error and concept possession and use, are at a more fundamental level historical and psychological facts. They are going to be facts about what physically goes on in the mind over time. This way of thinking about the theories in the Historical-Psychological family would make them trivially true. This sort of trivially true explanation is not the sort that is relevant for us.

What is relevant for us, are explanations that argue for some *particular* historical and psychological facts that when tokened differently produce conflicting political judgements. Of course, logically, there are many types of theories that could offer explanations like these. After all there are many types of particular historical and psychological facts that could plausibly be connected to moral and political judgement making. But, I take it to be fairly plausible that trying to find explanations that cite *particular* historical and psychological facts as difference makers, lends itself to restricting one's view to the type of theories best supported by empirical moral psychology. As we will see, this is precisely the type of theory I consider as part of the Historical-Psychological family.

In sum, with the two enrichments in hand – the idea of explanation and the expanded taxonomy – what follows in the next three sections are arguments against the

Imperfection Family, the Historical-Psychological Family and in favour of a particular version of the Conceptual Family, Concept Possession and Use theories.

3 Imperfection Family

I've said that the Imperfection Family can be summarised as the family of theories that argue facts about intellectual error are what make the difference between reasonable disagreement and reasonable agreement. These facts about intellectual error generally fall into two categories. They can be about a defect in reasoning, or about the use of an incorrect type of reasoning for moral and political matters. Theories that cite the first category of facts I'll call Defective Reasoning theories. Theories that cite the second category of facts, I'll call Wrong Type of Reasoning theories.

In §§3.1–3.2 I argue that both types of theories are inadequate explanations of reasonable disagreement. The Defective Reasoning theories rule out reasonable disagreement altogether. The Wrong Type of Reasoning theories either cite facts that do not make a difference between disagreement and agreement, or fail to justify the normative standard that would pick out such a fact. As a result, I conclude that no theory in the Imperfection Family is an inadequate explanation.

3.1 Defective Reasoning

I said earlier Defective Reasoning theories cite facts about defects in reasoning as the intellectual error that makes the difference between reasonable disagreement and reasonable agreement. The most obvious and I think most common, way to do this involves the familiar idea that disagreement arises because at least one individual is ignorant about some pertinent fact, or infers incorrectly when reasoning and forming a judgement. This is a familiar idea because it is a type of explanans we often give for disagreements that are not about moral and political matters. Disagreements in the natural sciences, economics, history, or political science are usually thought to arise by some form of ignorance or reasoning mistake which each side is trying to identify.

Of course there are different ways to cash out this general idea. One could cash it out epistemically. This would involve citing facts about people's ignorance of moral facts, the correct principles of rationality, or mistakes in applying such principles, as the intellectual error that causes disagreement.[3] The idea would be that reasoning

[3] See David Enoch (2011: 186–197, 207–214), Russ Schafer-Landau (2003: Ch.9), David Brink (1989: 197–210), David Wiggins (2006: 366–367) and John McDowell (1998: 162) for this sort of explanation. See Dworkin (2011: 441–446) for an overview of how the latter two's views may be construed this way.

about moral and political matters is like reasoning about any other matter. As such, getting the right answer requires attending to the moral facts and using the correct principles for rational thought. When one is ignorant of those facts or applies the principles incorrectly one is likely to arrive at any one of the whole range of wrong answers. At least one person doing this is what causes disagreement.

Another way one could cash out the general idea of intellectual error is linguistically. This would involve citing facts about people's ignorance of the meaning of moral terms, or mistakes in inferring from the meaning of one moral term to another, as the intellectual error that causes disagreement.[4] The idea here would be that moral and political terms have meaning and behave linguistically like any other terms. As such, using moral and political terms to make moral and political judgements requires attending to facts about what moral words mean and using them correctly given their meaning. When at least one person is ignorant of what their words actually mean or make mistakes in using them according to what they mean they will end up in a disagreement.

But, no matter how the general idea is cashed out Defective Reasoning theories cannot actually explain reasonable disagreement because they face a dilemma about its very existence. On one horn, Defective Reasoning theories entail that reasonable disagreement is not actually intractable understood in the sense of a disagreement persisting despite parties explaining their reasons for their judgements. This is because they could, according to Defective Reasoning theories, simply reason longer and acquaint themselves better with the relevant facts. After all what causes disagreement is not reasoning well enough or being ignorant of the moral facts.

On the other horn, if Defective Reasoning theories maintain that reasonable disagreements are in fact intractable, then they are committed to saying that the people involved are not really being sincere in their judgement making. This is because the only way their disagreement could be intractable given they could choose to reason longer and acquaint themselves with the relevant facts, is if reasonable people are making some wilful fault in their inferences or being wilfully ignorant of the relevant facts. But, part of what was settled as the explanandum that needs to be explained in the last chapter was that reasonable disagreement about justice involves intractably conflicting judgements made by people who sincerely want to make judgements that they think others can agree it. It is not plausible to think that people who are sincere in this way are wilfully ignorant of the relevant facts or wilfully making mistakes in

[4] See Andrew Mason (1993: 72–75) on how a Locke's (2008: 307–309, 314–315, 322–327) theory of language could be used to formulate such an explanation.

reasoning.

Given the dilemma, I believe it is safe to leave aside Defective Reasoning theories. No matter how they are cashed out they rule out one of the properties that constitute reasonable disagreement. To that end, I submit, we reject the Defective Reasoning theories and consider a more plausible type of theory within the Imperfection Family.

3.2 Wrong Type of Reasoning

Another type of theory within the Imperfection Family involves citing facts about the use of the wrong type of reasoning for making moral and political judgements. These are what I call Wrong Type of Reasoning theories which R.M Hare (1981) and Joshua Greene (2013) propose. These theories avoid the problem faced by the Defective Reasoning theories by simply denying that some fact about a defect in at least one party's reasoning is what makes the difference between disagreement and agreement. As such it does not face the dilemma that Defective Reasoning theories do. Rather, Wrong Type of Reasoning theories rely on the idea that reasonable disagreement about justice occurs because individuals choose a type of reasoning that is unsuited to the task of making political judgements about what justice requires.

Wrong Type of Reasoning theories begin by making a distinction between the type of reasoning suited to moral and political matters – "critical thinking" for Hare (1981: 40) and "manual mode reasoning" for Greene (2013: 133-134) – and the type of reasoning that causes disagreement. Hare (1981: 40) describes this distinction and the motivation for it when he says:

> ...however well equipped we are with these relatively simple, prima facie, intuitive principles or dispositions, we are bound to find ourselves in situations in which they conflict and in which, therefore, some other, non-intuitive kind of thinking is called for, to resolve the conflict...
>
> What will settle the question is a type of thinking which makes no appeal to intuitions other than linguistic. I stress that in this other kind of thinking, which I am calling critical thinking, no moral intuitions of substance can be appealed to.

For Hare, explaining why reasonable disagreement occurs amounts to picking out the type of reasoning about moral and political matters which doesn't lead to conflict. For Hare (1981: 40–42, 153–160), intuitions are unsuited for this task because people will inevitably differ in their intuitions because of their different "upbringings and past experience". Using different intuitions either directly, or indirectly by constructing

what Hare (1981: 40-41) calls "prima facie principles", to make judgements will result in conflicting political judgements. We are not justified in picking one intuition over another because it is an open and substantive question which kind of upbringing or set of experiences are the correct ones to have. The conflict between two such judgements persists by trying to arbitrate between them using further intuitions, when we ought to use "critical thinking".

This is a type of reasoning that seeks to universalise our judgements from the situation being judged to "any of the other precisely similar situations". According to Hare (1981: 42) what this sort of universalizing amounts to is the following:

> What critical thinking has to do is to find a moral judgement which the thinker is prepared to make about this conflict-situation and is also prepared to make about all the other similar situations. Since these will include situations in which he occupies, respectively, the positions of all the other parties in the actual situation, no judgement will be acceptable to him which does not do the best, all in all, for all the parties.

The core idea here is that making the correct choice in the type of reasoning to use – critical thinking – will involve finding moral and political judgements that universalise across contexts and people's particular positions in those contexts. Doing this will yield judgements that "do the best, all in all, for all the parties". Not engaging in this sort of reasoning is what leads to reasonable disagreement. This is because it will involve making judgements that are not acceptable to all parties because they will be picking out what is morally significant for only some individuals in some contexts.

Unlike Hare, Greene (2013: 14–15, 23, 26) begins with a distinction between two kinds of disagreements: first-order moral disagreements and second-order moral disagreements. First-order moral disagreements are disagreements between those who share our basic moral intuitions about what individuals ought to do. Second-order moral disagreements, are disagreements about what our society ought to do with those who do not share our basic moral intuitions about what individuals ought to do. Greene argues that using moral intuitions to reason is helpful for resolving the first kind of disagreement, but not the second kind. Using intuitions works for the first kind because we are disagreeing with those who share our intuitions and resolving a disagreement amounts to clarifying our intuitions or adjusting them to changing circumstances. But, these strategies do not work for the second-order moral disagreements because they involve disagreements about what is intuitively of value or normatively required.

Greene introduces, with empirical justification, a model for how people reason

about moral and political issues that is supposed to explain why moral intuitions do not help resolve the second-order disagreements and what kind of reasoning does help us. Greene (2013: 117, 2014a: 696–707) proposes the "dual-process" theory of moral judgement making. The core idea being that human beings have two ways of reasoning to make judgements. On the one hand we can use an "automatic" mode of reasoning that involves making intuitive judgements. Even though we might then rationalise these judgements using deliberative reasoning, for Greene, their foundation is in intuition or emotion. On the other hand we can use a "manual" mode of reasoning which involves making judgements by considering what is of fundamental value, abstractly and counterfactually in a conscious and explicit way. Roughly, Greene (2013: 120, 137–141) believes automatic and manual mode reasoning correspond to how we ought to be forming judgements when faced with two different kinds of decisions. When making decisions with those who share our intuitions we ought to use the automatic mode, but when faced with those who do not share our intuitions we ought to use the manual mode.

According to Greene (2013: 54, 62–63, 2014a: 698), humans use the automatic mode of intuitive reasoning because we are evolutionarily "hardwired" to cooperate. This disposition means that individuals favour a style of reasoning that is quick and efficient for judging others and situations relevant for the cooperative success of the group. However, this style of reasoning is only successful in small communities where for the sake of cooperation individuals tend to conform and share each other's intuitions. But, when faced with situations where communities that do not share intuitions must reason together, reasoning with intuitions leads to disagreement. What causes these disagreements is a state of affairs where individuals choose the "automatic" mode of reasoning. This causes individuals with different moral intuitions to form conflicting moral and political judgements.

In sum, Wrong Type of Reasoning theories like Hare and Greene's explain reasonable disagreement about justice by citing the fact that reasonable people choose the wrong type of reasoning, namely a type of reasoning that uses intuitions, to make their moral and political judgements. Explaining reasonable disagreement about justice in this way also avoids the problem that faced Defective Reasoning theories.

Hare and Greene do not rule out disagreement between reasonable people. The kind of intellectual error Hare and Greene's theories cite does not involve saying that at least one of the disagreeing parties has made a factual or inferential mistake. Rather Hare and Greene only say that there is a correct view about the sort of reasoning reasonable people ought to use to make moral and political judgements when they

are trying to live in a shared social world. It is an intellectual error in this choice of reasoning that causes reasonable disagreement about justice.

Hare and Greene's theories also do not rule out reasonable disagreement involving intractably conflicting judgements. On their theories it is perfectly understandable why this occurs. As Greene (2013: 62–63, 2014b: 1018) argues, individuals from different communities continue using their respective intuitions to reason because doing so has been successful in the past. Likewise, Hare (1981: 39) argues that moral and political judgements formed by reasoning at the intuitive level are "necessary for human moral thinking". This is because intuitive reasoning yields simple general principles from which we can learn how to make moral and political judgements quickly for a great variety of situations. In short, it is successful in helping us make personal decisions and also collective decisions with others that largely share our intuitions and sentiments. The versatility of using intuitive reasoning means that reasonable people are not motivated to suddenly switch to a type of reasoning that would allow them to agree with each other.

But, I submit, Hare and Greene's theories face a number of problems when we look closer at their claims about what precisely makes one type of reasoning erroneous and another the correct for making moral and political judgements.[5] For Hare (1981: 39–41) this comes down to the idea of universalisability. Reasoning that aims to make judgements that universalise across similar situations and regardless of the judgement makers role in those situations will be the correct type of reasoning. For Greene (2013: 117) the relevant idea is that manual mode reasoning involves the capacity to consider what is of fundamental value, abstractly and counterfactually. This allows people to put aside their conflicting intuitions. For Hare (1981: 111–115) and Greene (2013: 171–174) the judgements that will tend to emerge from their the correct type of reasoning are utilitarian ones or judgements that pass the utilitarian normative standard for right action or rules.

On Hare and Greene's theories then, reasonable people can avoid disagreeing only if they make the correct choice when reasoning about political matters. That is, only if they set aside, or at least limit the influence of, moral intuitions and emotions. But, even on Hare and Greene's own examples where their respective explanations are supposed to work there is still an open question whether the difference makers they cite do cause disagreement rather than agreement. This is because in the case of Hare many moral and political theories claim to offer judgements that satisfy universalis-

[5] See Kumar and May (2018: 34–39) for a version of this criticism target specifically at Greene's theory.

ability. In short, universalisability is a too minimal, and too widely shared, idea to really select some type of reasoning as a cause of disagreement. Disagreement occurs even with the correct choice of reasoning. In Greene's case, the difference makers fail to show that either, instances of manual mode reasoning will involve deliberations of the same sort and the use of the same values, or that some amount of autonomic intuitive reasoning will not be present. In short, the difference makers do not actually establish that choosing the correct type of reasoning can avoid disagreement.[6]

For instance, Hare (1981: 155) considers the example of a disagreement involving the just administration of the right to free speech. In the example, a "public authority" refuses an openly racist organisation from reserving the use of a public hall. The disagreement emerges when, as Hare (1981: 155) says:

> The racist organization then protests that it is being denied its right to free speech. The public authority counters that it has an obligation to preserve the right of minorities not to have hatred preached against them, and that the public has a right to be protected against outbreaks of violence.

For Hare what makes the difference between disagreement and agreement in the case above is that at least one of the parties makes their normative judgement by appealing to, or with the justification of, intuitions. Specifically that the racist organisation and the public authority will merely justify their judgements about the free speech rights justice requires by citing further intuitions. What the parties ought to do, according to Hare (1981: 156), is step back and recognise that the only way to avoid disagreement is to apply a mode of critical thinking which weighs up what ought to be done according to the "logical considerations established by an understanding of the words used in the questions we are asking, and compelling on anybody who is using the words in the same senses". For Hare (1981: 156) this means deciding what is just to do because it conforms to principles "on the ground that they are the set of principles whose

[6] Of course there is another way to criticise Greene's theory which involves criticising his epistemological and empirical claims. This is to criticise the distinction between intuition based judgement making and abstract, counterfactual, critical reflection based judgement making. This amounts to showing that their evaluation of the former is unreliable or epistemically suspect in some way compared to the latter. See Prinz (2016) and Driver (2016) for arguments of this kind. I avoid this sort of criticism for two reasons. The first reason is that much more empirical work needs to be done to decide whether intuition based judgement making is completely separate from any form of reasoning or deliberation. The second reason is that such arguments do not target the core claim of the Wrong Type of Reasoning version which is that there is a fact of the matter about what type of reasoning one ought to use in deliberative about moral and political matters and that the wrong choice causes disagreements.

general acceptance in the society in question will do the best, all told, for the interests of the people in the society considered impartially". The idea is that adopting a type of reasoning that seeks to universalise our moral judgement making in the way utilitarianism asks of us will lead to agreement. A failure to do this will lead to disagreement. In the case above, Hare believes that using critical thinking and seeking to universalise our judgements will lead us to judge that the public authority ought to guarantee broad freedom of speech qualified by restrictions on what specific things may be said. This best approximates what, according to Hare, the utilitarian would judge.

The problem with all this is that many mutually incompatible theories about justice propose, or at least aim to propose, normative standards that satisfy the idea of universalisability. Kantians, virtue ethicists, contractualists or any variety of non-Kantian non-consequentialist deontology can all propose normative standards that conform to the idea of universalisability.[7] For instance in the case of the racist organisation, libertarians will counter that their judgement about the free speech rights stems from it being a natural right and the principle that individuals have a natural moral right to live and pursue projects as long they do not harm an other's right to do the same. And, they have come to *this* principle by thinking about the sort of rights all individuals have regardless of who they are or their projects. As such, they will claim that the right to free speech is something that all individuals have where as no individuals have the right to not be subject to hate speech. The point here is that non-utilitarians can equally claim to be following normative standards that satisfy universalisability. This leaves it open for two individuals to use critical thinking and yet continue to disagree because the political judgements they make can conform to the universalising of different normative standards. Both the libertarian and Hare's utilitarian can claim to use a type of reasoning that looks to universalise their moral judgement making. It is merely that they both disagree about the normative standard that correctly does this job.

Given all that, the use of intuitive reasoning is not really what makes the difference between reasonable disagreement and reasonable agreement. After all, the sort of reasoning that was supposed to make the difference by satisfying universalisability – critical thinking – can also lead to disagreement. This means that Hare does not actually tell us why reasonable disagreements occurs rather than reasonable agreements. Rather Hare only contrasts two ways that disagreement can arise. He does not tell us why the wrong choice of making judgements with intuitive reasoning leads to reason-

[7] See George Sher (1984: 183–184) for similar criticism.

able disagreement, but critical thinking does not.

This is similar to Greene's example of abortion. Greene (2013: 309, 322) uses abortion as a case study for his explanation by taking the real example of disagreement between "pro-choice" and "pro-life" judgements in the United States of America. What interests us is whether Greene's difference maker – manual mode reasoning – actually makes the difference between disagreement and agreement. Greene strips the disagreement down to what he considers to be two honest positions. This means stripping the positions of judgements that appeal to phenomena the other side does not accept, like souls or "confident talk about a 'right to life' and a 'right to choose"'. Greene (2013: 321–322) then summarises the disagreement as a clash of political judgements formed by the use of moral intuitions about the innocence of human souls and the coercion of women. Greene summarises the "pro-life" position as:

> You can't rightly kill an innocent human soul. I know that this is partly a matter of faith. And I understand that we're supposed to respect each other's beliefs. Bur I just can't see letting people kill something, even if it's small, so long as there might be a human soul in there. I know that's hard on a lot of people who don't want to be pregnant. But, those people made a choice to have sex (except in the case of rape, which is different), and killing something that maybe has a human soul is not a legitimate way to undo that choice. That's how I feel.

He then summarises the "pro-choice" position as:

> Forcing a woman to do that seems worse to me than killing a froggy little human. Third trimester fetuses, however, don't look froggy. They look like babies. And killing babies is clearly wrong. So if the fetus you're carrying looks kinda froggy, then I think it's okay for you to kill it, if that's your choice. But, if your fetus looks like a real baby, and not a little froggy thing, then I think you have to let it live, even if you don't want to. That's how I feel.

According to Greene, using manual mode reasoning in this case amounts to weighing up the consequences of acting on each of the opposing political judgements and choosing the one that brings about the best consequences measured against a common currency of human values. Which currency? According to Greene (2016: 175, 2013: 161), the common currency of human values is happiness, or qualitatively positive experiences. The justification for this currency is that it is what all human beings intrinsically value and so it is impartial with respect to a disagreeing party's intuitions

and it is general enough to use in manual mode reasoning in a range of disagreements regardless of the specific judgements in play. As Greene (2016: 175) says:

> ...I argue that deep pragmatism (utilitarianism, properly understood and wisely applied) is our best bet for a global "metamorality," a higher-order normative standard that adjudicates among competing tribal values and interests, just as a single tribe's value system adjudicates among the competing values and interests of its individual members. I do not claim that deep pragmatism is the moral truth, only that it's our most promising metamorality.

The point of all this is that given Greene's theory, when we take happiness or qualitatively positive experiences as the common currency of human values and combine it with the account given of manual and automatic reasoning, we get an explanation of why reasonable disagreement occurs and how it can be avoided.

When applied to the example of abortion, the consequences of the "pro-choice" judgements are that forcing women into pregnancies can, when the woman does not want to have the pregnancy, place great emotional strain and perhaps lead her to care less for the new born child. All round the consequences of legalising abortion is that it prevents having to forcing women to do something against their will that create lots of unhappiness. For Greene, once we leave aside the pro-life judgements based on metaphysical beliefs about the soul which pro-choicer's do not share, we can weigh the consequences of the "pro-choice" judgement in purely manual mode reasoning. Pro-life judgements would lead to more people existing and would most likely result in a net gain in happiness. According to Greene the choice of manual mode reasoning leads us to accept the "pro-choice" judgement. This is because it is too much to ask of "nonheroic people" that they weigh their own happiness less than the happiness of non-existent lives with the potential to be happy. To that end Greene (2013: 325–326) thinks more happiness is created by arbitrarily drawing a line between when a human can be killed in the womb and when it can't rather than "Disrupting people's sex lives, disrupting people's life plans, and forcing people to seek international or illegal abortions". This is the conclusion Greene thinks each party would reach if they used pure manual mode reasoning. Reasonable disagreements about the laws relating to abortion would be avoided. It is because individuals choose not to reason about issues using manual mode reasoning that disagreement arises and persists. Using the automatic mode of reasoning by appealing to moral intuitions and parochial values is what causes reasonable people to disagree.

There are, I submit, three problems with Greene's explanation. The first problem

is that the explanation does not actually justify why manual mode reasoning would entail the sort of calculation about demandingness that Greene supposes in the abortion case. The second problem is that it does not establish that manual mode reasoning, at least in the abortion case, relies on intuitions any less than automatic mode reasoning. The third problem is that Greene's explanation does not establish that happiness is the common currency of reasonable people's shared moral values.

The first problem emerges because Greene's argument for why manual mode reasoning (and not automatic mode reasoning) resolves a disagreement like the abortion case is that it hinges on a particular view about the calculation that would be involved in using manual mode reasoning. Specifically the calculation is that it is too demanding to ask of living people to prefer to bring about more potentially happy people through outlawing abortion than to prefer the happiness of living people who do want to get an abortion. Greene refers to this demandingness as consequences that are "*too good*" because it demands restrictions on other aspects of life like contraception and abstinence. But, on the other hand as Greene (2013: 325) says "the pro-choicer's utilitarian arguments are not *too* good. They're just plain good". They appropriately weigh up the consequences of "Disrupting people's sex lives, disrupting people's life plans, and forcing people to seek international or illegal abortions". Simply put manual mode reasoning would lead us to the pro-choicer's judgement because of how that type of reasoning forces us to accurately weigh the consequences of abortion.

But it seems entirely consistent that someone might make the very opposite weighing of the consequences of the pro-life or pro-choice positions when using manual mode reasoning. This is not philosophical speculation. Kahane et al. (2012) have shown that that manual mode judgement making cuts across non-utilitarian judgements. In short, people do make non-utilitarian judgements when using their manual mode reasoning. Greene does not offer any principled reason why the moral demandingness of the pro-life judgement's consequences are decisive in not taking it seriously. This is a straightforward case of either probabilities that are vague or indeterminate because the consequences involve humans who do not exist yet. Greene's judgement that the pro-life judgement is too demanding assumes a particular view about how potentially happy non-existent people will be. But, a range of plausible conflicting judgements appear possible when guessing how non-existent people will turn out or the happiness they will derive from their experiences. For instance the pro-lifer could point to the equally demanding consequences for women who want to become pregnant and yet live in a society that assumes the rational course of action is to have an abortion given their financial situation. The pro-life could also temper Greene's ar-

gument by point to the possibility of people who have abortions who would have had happier lives if they had been prevented from doing so. The pro-lifer could argue that a society in which not having abortions becomes a norm is one where society will see the need to improve the financial situation of those with children. This would definitely not have the demanding consequences that Greene supposes. The point of all this is simply that much more needs to be said to establish that manual mode reasoning would yield the sort of calculations about demandingness that entail the pro-choicer's judgement about abortion rights.

The second problem with Greene's explanation arises because even if we assume that the calculations involved in manual mode reasoning can be shown to resolve reasonable disagreements, it is not clear why it would not involve the use of moral intuitions. For instance, Jesse Prinz (2016: 57–60) argues experimental evidence shows that the intuitions that characterise automatic mode reasoning are equally present in cases where people make the utilitarian judgements that characterise manual mode reasoning. He argues the more plausible interpretation of Greene's studies is that some people regulate their emotions in different ways no matter how they make their moral and political judgements.

Also, James Woodward (2016: 87–93, 104–105) argues experimental evidence shows that the neural structures that Greene classifies as used for automatic mode reasoning – ventromedial prefrontal cortex, orbital frontal cortex, anterior cingulate cortex, insula and amygdala – are actually all used in information processing and that Greene's view relies on the mistaken assumption that evaluative and practical judgements could be made without those neutral structures. Woodward argues that these neural structures are crucial in processing uncertain consequences and any sort of value like human happiness. The point here is that when making judgements with the sort of utilitarian calculation that Greene says will avoid disagreement we do not know the probability distributions on the various effects of our actions. This means that it is not obvious that judgements made using manual mode reasoning will be much different from the judgements reached by automatic mode reasoning. This is because the range and depth of information that needs to be considered when using manual mode reasoning consistently in political disagreements will involve estimates which will be open to the same kinds of differences as the intuitive differences in automatic mode reasoning. Woodward argues that when people use the sort of manual mode reasoning that Greene thinks will avoid disagreements like in the abortion case, they will really be making complex calculations that involve reasoning about how best to calculate and increase happiness given all the information they have about the past

and the future. This sort of reasoning is not merely tallying happiness or unhappiness, but also calculating the best strategy to increase happiness in the long-term. Woodward thinks this sort of complex calculation will at best allow and at worst require making judgements using one's emotions or intuitions. If this is the case then manual mode reasoning is as likely to lead to disagreements as automatic mode reasoning. Given all this, Greene's explanation of the abortion case does not hold. As an example of manual mode reasoning it shows reasonable disagreement can arise and persist despite the use of such reasoning.

The third, and most worrying problem for Greene's explanation is that even if we grant that the calculations involved in manual mode reasoning can be largely done without the presence of moral intuitions, he provides no justification for happiness as the currency of those calculations. Whilst Greene's justification for manual mode reasoning as resulting in broadly consequentialist normative judgements is supported by considerable empirical evidence, the theory of value that Greene needs to justify manual mode reasoning as the right type of reasoning is not supported by the same evidence. Greene's (2013: 190–192) only justification for happiness as the common currency in our shared moral values is a series of rhetorical questions about why we "care" about certain states of affairs. Greene's conclusion is that all our desires and concerns are grounded in improving the quality of experiences for ourselves or others and decreasing the harm and displeasure. But, this is a conjecture. It is entirely consistent with the "pro-life" position that those who hold it do not take happiness as what grounds all our desires and concerns. In fact we might think that part of what a reasonable disagreement concerning abortion is really about is precisely what in the consequences of our judgements we ought to care about. Greene gives us no reason why those who endorse a ground other than happiness are being any more irrational or incoherent than those who endorse happiness as the ground.[8] Greene cannot justify citing the difference maker he does without giving us such a reason. For Greene's theory to work he needs to justify the causal claim that it is only the intuitive automotive mode of judgement making that makes the difference in bringing about a state of affairs of disagreement rather than agreement. But, this is precisely what Greene cannot do without assuming happiness as the normative standard for when political judgements are erroneous. It is consistent with Greene's theory that two individuals each of whom uses manual mode reasoning but who take different things as the 'com-

[8]See Kahane (2016: 291–292) for a similar point but in relation to treating Greene's explanation as a debunking explanation of the reliability of intuition based judgement making. Kahane argues that characterising intuitions as less reliable ways of making true judgements fails without assuming utilitarianism as the normative standard for judgements.

mon currency' of all our desires and concerns will result in a reasonable disagreement. Greene's theory needs to go beyond conjecture and until his theory does so it cannot cite the difference maker it aims to.

Given all these problems, I submit, it is safe to reject the Wrong Type of Reasoning theories. I argued in various ways against both Hare and Greene that, either they cannot actually justify why a particular type of reasoning ensures reasonable disagreement will not occur, or they cannot establish a clear enough distinction between the correct type of reasoning and the incorrect type of reasoning, or they do not establish the normative standard that distinguishes the correct type of reasoning will not itself be the subject reasonable disagreement. Taken collectively I believe this gives us good reason to reject the Wrong Type of Reasoning theories and as a consequence reject the Imperfection Family.

4 Historical-Psychological Family

In the last section I argued that we ought to reject the Imperfection Family of theories for explaining reasonable disagreement on the basis that both of the types of theories in that family fail. The Defective Reasoning theories either rule out reasonable disagreement about justice altogether, and the Wrong Type of Reasoning theories either cite a fact that does not make a difference between disagreement and agreement, or fail to justify the normative standard that would pick out such a fact.

One obvious way to try and avoid these problems is to leave aside the notion of intellectual error that characterises the Imperfection Family of theories. Rather than looking for difference makers in facts about the sort of intellectual error that can occur in moral and political judgement making, we might think a better explanatory strategy is to look for difference makers in facts about the psychology and histories of the judgement makers and how these facts affect what judgements they make. Theories like this comprise the Historical-Psychological Family.

As I have already mentioned, the strategy of looking to particular facts about the psychology and histories of the judgement makers to explain reasonable disagreement, lends itself to the type of theories best supported by empirical moral psychology. The most developed contemporary theory like this is Jonathan Haidt's Moral Foundations Theory. It proposes differences in moral intuitions and constructed life narratives as the psychological and historical facts that make the difference between reasonable disagreement and reasonable agreement.

This might seem odd. After all, there are other theories of moral psychology

about the nature of moral judgement making. For instance, there is a long-tradition in psychology from Piaget and Kohlberg, to Turiel, and more recently Nichols on the nature of moral development and the capacity for moral judgement making.[9] Why not focus on these? I offer two reasons.

The first reason is that, as I just said, they have so far been concerned with providing empirically grounded accounts of people's capacity for moral judgement making. They have not focused on explaining psychological differences in judgement makers as a way to explain the type of moral and political disagreement between reasonable people. This is the core phenomenon in reasonable disagreement about justice. As such, they simply have not targeted the explanandum I am considering in this thesis. In contrast, Haidt's (2012: 9) Moral Foundations Theory aims to explain as he says, "...why it's so hard for us to get along...why we are so easily divided into hostile groups, each one certain of its righteousness". Haidt's focus on the "righteous mind" is a focus on a state of affairs where each participant in a reasonable disagreement insists their judgement is correct despite interaction with other parties. This focus is also borne out in the examples that Haidt uses. Particularly the ones between the major political parties in the USA. Haidt's focus in all this is precisely at what leads reasonable people (ie. those with a minimally adequate level of moral development) to form into distinct disagreeing groups within and across cultures.

The second reason for not focusing on the long tradition of moral psychologists is that the little that these theories do have to say about disagreement largely involves, in the case of Piaget (1965) and Kohlberg (1984) citing a lack of moral development, and in the case of Turiel (2002) and Nichols (2004) citing differences in intuitions and information assumptions about their social world, that produce different core normative theories that people use to make judgements. In Piaget and Kohlberg's case the explanation collapses into a type of theory in the Imperfection Family (which we already rejected). In Turiel and Nichols's case the explanation ends up being the sort of historical-psychological explanation I argued was too general to be relevant for our purposes. This of course doesn't mean that they are not relevant for explaining reasonable disagreement whatsoever. As I'll show later in this section, they provide compelling evidence against aspects of Haidt's theory and show how we need to go beyond moral psychology for an adequate explanation of reasonable disagreement.

With all that in mind, the rest of this section proceeds as follows. In §4.1 describe Haidt's theory and detail how it proposes to explain reasonable disagreement. In §4.2

[9] There are of course more theorists in the tradition but I mention the most prominent touchstones for philosophy.

I argue that it cannot do what it aims to do and how as a result we have good reason to reject the view. I will argue that the theory fails on the measure of explanatory power because it cannot account for disagreements that result from people changing their judgements through reflection. The theory both runs contrary to the obvious fact that people do change their minds through reflection and relies on a model of moral cognition that is itself empirically contested. Both of these problems mean it offers an implausible explanation of reasonable disagreement about justice.

4.1 Moral Foundations Theory

For Haidt, the facts that make the difference between reasonable disagreement and reasonable agreement are the differences in reasonable people's moral intuitions, or more specifically the pattern of moral intuition in their cognitive process of making moral and political judgements (ie. their moral cognition). Haidt (2012: 147, 194–195, 198) argues these differences in a person's moral intuitions are modelled as "moral matrices" within the framework of six foundational dimensions of morality. Each dimension corresponds to a particular intuitively desired or hated type of personal and social activity. This framework of the foundational dimensions of morality is a result of the kinds of personal and social actions that were naturally selected for. Haidt (2012: 94–96, 158–160, 214) cities anthropological studies as the primary evidence that the framework exists across cultures and particular human communities. What determines the particular moral matrices that particular individuals have on this framework, are the intuitions people are born with and childhood experiences that reinforce these intuitions.

An individual's moral and political judgements are caused by the combination of genes which determine a brain structure with a pattern of moral intuitions and then childhood experiences that reinforce particular intuitions over others (Haidt 2012: 166, 328–336). This is why the moral matrix that leads a person to make particular moral and political judgements persist despite interaction with those they disagree with. The unalterable causal influence of one's genes and childhood experiences continue to determine one's brain structure into adulthood. This explains the evidence that Haidt (2012: 118–119) points to where adults experience disagreement that persists despite deliberation with other interlocutors. Making sense of or explaining their childhood experiences – understood by Haidt as the 'construction of life narratives' – solidifies an individual's particular moral matrix for what she considers to be morally right or just. This is because as Haidt (2012: 334) says human beings construct narratives as part of a process to confirm their identity within their community. Indi-

viduals see their childhood experiences of being taught a preference for certain moral sentiments as part of their intrinsic identity. For Haidt the construction of life narratives is a psychological fact about how human beings reflect on their childhood experiences. This fact of life narrative construction "binds and blinds" by which Haidt (2012: 366) means the life narratives reinforce the determining forces of one's genes and childhood experiences.

For Haidt (2012: 42–45, 121), all of this theory so far is situated in a particular dialectic where he is arguing against what he calls the "Rationalist". He dubs the "rationalist delusion" as the view that abstract deliberative reasoning – understood as the weighing of evidence, desires, norms and the reasons they all generate – plays the central role in determining the moral and political judgements individuals form. Haidt's (2012: 27, 34–36, 41–42) main evidence against the "rationalist delusion" are the experiments he cites that show children and adults across cultures make moral and political judgements without abstract reflective reasoning. Haidt argues that this points to something other than reasoning playing the central role in generating moral and political judgements. In addition, Haidt presents evidence that shows what appears to look like reflection when making moral and political judgements is actually post-hoc rationalising which involves looking for reasons for a judgement that has already been made. Haidt (2012: 43–45, 59–63) cites evidence that children invent harms and victims of harming in the face of questioning about their moral and political judgements. With such evidence Haidt argues that the Rationalist view is unsupported by modern theories of moral psychology. Rather the anthropological, psychological and neuroscientific evidence he cites supports his theory of moral psychology and the difference makers that it cites as the facts that cause reasonable disagreement.

All of this comes together in what Haidt (2012: 71) calls the "Social Intuitionist Model" of moral judgement making. On this model, moral judgement making begins with moral intuitions which are a form of cognition. The particular set of moral intuitions an individual has and is disposed to employ make up their moral matrix. This matrix is determined by her genes and personal childhood experiences. The firing of a particular moral intuition in response to some personal or social activity, then causes reasonable people to make moral and political judgements. These judgements are either reinforced by other people's intuitions, judgements and reasoning or reinforced by our post-hoc reasoning. In both cases an individual forms a narrative that confirms certain parts of their moral matrix over others. From there the cycle begins again with reasonable people's moral intuitions firing and causing moral and political judgements. In sum, when an individual's genes and personal history diverge from

her interlocutor a disagreement is caused. This is because these are the facts that cause reasonable people to have different moral matricies and different moral matricies between reasonable people is the fact that makes the difference between reasonable disagreement and reasonable agreement.

All this immediately improves on the Imperfection Family of theories in two ways. The first way is that, Haidt's theory does not rule out the very phenomenon being explained. It does not rule out disagreements being intractable or between reasonable people. The second way is that it avoids having to ground the difference makers in any property of reasoning, substantive normative standard or any facts that might itself be the subject of reasonable disagreement. It only needs to cite differences in a restricted set of facts about a reasonable person's psychology and history as difference makers which might in turn ground various things about human reasoning.

4.2 Reflective Judgement Making

There are, I submit, two problems with Haidt's theory. Both of which arise because of how he responds to cases of reflective judgement making. The first problem is that Haidt's response – that such cases are rare if they occur at all – is not supported by the empirical evidence on how people make moral and political judgements. The second problem is that Haidt's response, and broader theory, relies on an implausible model of cognition for moral judgement making because it also runs contrary to lots of empirical evidence. I believe both problems give us good reason to reject Haidt's theory, and accept a more modest view of the psychological basis of moral and political judgement making.

The two problems with Haidt's theory emerge when we consider cases where reasonable people make new moral or political judgements after reflecting on their previous judgements. This sort of reflective judgement making can occur in two ways. One way is that reasonable people in reflecting on their previous judgements, reconsider how much weight or how much emphasis they placed on one part of their moral matrix over another. This can then lead them to rerun their judgement making in a more deliberative way and make a new different judgement. A clear example of this is when people hear new arguments or have new experiences and weigh up deliberative considerations differently.

Another way reflective judgement making can occur is when reasonable people reflect on the place of deliberative considerations in their reasoning itself. After hearing entirely new arguments, having new transformative experiences or by simply questioning why they have certain intuitions or take certain considerations seriously they

may drop a consideration altogether or adopt new ones. This then leads to a new deliberative process and making a different judgement. A clear example of this would be when people ask themselves why they have certain long held assumptions. They may find they can't see them as plausible any longer and discharge them or replace them with new assumptions.

In both ways that reflective judgement making can occur reasonable people can find themselves in disagreement with new interlocutors or in renewed disagreement with old interlocutors. That this occurs is obvious and beyond doubt. Whilst it might be plausible to imagine reflection plays no role in any initial judgement making in children or adolescents, it is obvious it plays a role when individuals mature and begin to think reflectively. From ordinary people to moral philosophers, people make new judgements through reflection and find themselves in new reasonable disagreements.

But Haidt's theory, I submit, does not have the ability to cite the facts that explain how those cases of reasonable disagreement arise. This is because the whole theory is based around showing how intuitions are the sole cognitive element that cause people to make moral judgements. Even when people make new judgements that conflict with earlier ones it is because their intuitions change according to other people's intuitions or judgements. This is a severe explanatory weakness. Reasonable disagreements obviously occur as a result of reflective judgement making. Not explaining such a case is worrying because Haidt's theory was supposed to be an improvement on the Imperfection Family of theories. But, theories in the Imperfection Family can easily explain cases of reflective judgement making. They will simply cite the facts about the intellectual error that individuals make in their process of reflection. If theories like Haidt's are supposed to be a serious improvement on the Imperfection Family they need to at least say something that explains cases of judgement change through reflection.

The two problems for Haidt's theory arise because of how he responds to this explanatory weakness. Haidt argues that reflective judgement making that results in new judgements is rare. Whilst reflection is a possibility, judgement change occurs most of the time, if not all the time, by the social influence of other people's judgements. Other people's judgements trigger new intuitions that can change judgements. But ultimately, moral and political judgements according to Haidt all start by the firing of moral intuitions. As Haidt (2012: 71) says:

> We make our first judgements rapidly, and we are dreadful at seeking out evidence that might disconfirm those initial judgments. Yet friends can do for us what we cannot do for ourselves: they can challenge us,

> giving us reasons and arguments (link 3) that sometimes trigger new intuitions, thereby making it possible for us to change our minds. We occasionally do this when mulling a problem by ourselves, suddenly seeing things in a new light or from a new perspective (to use two visual metaphors)....For most of us, it's not every day or even every month that we change our mind about a moral issue without any prompting from anyone else. Far more common than such private mind changing is social influence.

Unfortunately Haidt provides little argument or evidence to support his claim that judgement change through reflection is rare.[10] In fact, Haidt (2012: 71, fn. 44) acknowledges his lack of evidence, but does not see it as a problem when he says:

> One of the most common criticisms of the social intuitionist model from philosophers is that links 5 and 6, which I show as dotted lines, might in fact be much more frequent in daily life than I assert. See, for example, Greene, forthcoming. These critics present no evidence, but, in fairness, I have no evidence either as to the actual frequency in daily life with which people reason their way to counterintuitive conclusions (link 5) or change their minds during private reflection about moral matters (link 6). Of course people change their minds on moral issues, but I suspect that in most cases the cause of change was a new intuitively compelling experience (link 1), such as seeing a sonogram of a fetus, or an intuitively compelling argument made by another person (link 3). I also suspect that philosophers are able to override their initial intuitions more easily than can ordinary folk, based on findings by Kuhn (1991).

Haidt's position here is to maintain that making new judgements through reflection is rare and so any explanatory weakness is minimal. His main defence seems to be that the empirical evidence supporting his theory of moral psychology with its Social Intuitionist Model is compelling enough to show that most changes to earlier judgements occur through new intuitions. On this view reflective and unreflective judgement making involve distinct cognitive mechanisms. For Haidt, the latter is what overwhelming produces moral and political judgements. Any appearance of reflection on how one made one's political judgement is really post-hoc rationalisation

[10] Prinz (2016: 61–62) has the same worry, but goes further and says that it is unlikely Haidt *can* ever provide evidence or this sort because it would involve a "massive reservoir of field data" of varied people making moral judgements throughout their ordinary daily life.

that involves convincing one's self that we have good reason to make the judgement in the first place (Haidt 2012: 366). The point here is that the best explanation of reflective judgement making is that it is either rare or merely appears as such.

The first problem with Haidt's response is that regardless of the evidence in favour of the Social Intuitionist Model, the empirical evidence is not clear that reflective judgement making is rare. Empirical evidence shows that when people are either, primed by non-moral and non-political questioning that requires reflection to answer correctly, or given time to reflect on their judgements, they do tend to be reflective and reinforce or override the judgements they make based purely on initial intuitions.[11] Importantly this evidence is of a random sample of participants and not university students or philosophers who Haidt suspects do reflect and change judgements more frequently than others. Aside from this, Daniel Jacobson (2012: 298–304) has argued recently that much of Haidt's interpretation of the empirical evidence relies on an equivocation between whether people are unable to state a reason for their moral and political judgements and whether people do not have any reason at all for their moral and political judgements. It is evidence for the latter that would show that reflection in moral and political judgement making is rare. But, Haidt only presents evidence for the former. All this suggests that we have at least *some* reason to think reflective judgement making is not as rare as Haidt thinks.

The second problem with Haidt's response is that his model of moral cognition, is not as clearly supported by the evidence as he claims. Recently some have argued that a view that recognises deliberative reasoning and intuitions as both involved in moral cognition to produce moral and political judgements is a better view than Haidt's. For instance, Kennett and Gerrans (2016: 76–82) argue that most of the evidence Haidt cites as support for his model of moral cognition is based on individual responses to cases where they are forced to make moral decisions which are "synchronic" (for Kennett and Gerrans this means instantaneous verdicts unconnected over time). But, moral decisions are not made this way in our daily lives; moral decisions involve deliberation over time and are often embedded in seeing ourselves as "diachronic agents" where our actions over time are causally interconnected. This seems particularly relevant because moral and political judgements are the sorts of judgements that have multiple downstream effects that will give rise to making further judgements. Kennett and Gerrans note that the separation of intuition based judgement making and deliberative reasoning based judgement making is only supported *if* we view moral and political judgements in a purely synchronic way. When viewed as more complex

[11] See Stanley Et Al. (2019) and Paxton Et Al. (2011: 8–11).

cognitive acts we have reason to think that intuitions are not purely non-rational triggers for political judgements, but rather ways to deliberate about various social, causal and immediate pressures that will affect our actions. The use of intuition to make judgements contains reasoning of a sort.

Some have taken the further step to question Haidt's claims about intuitions as cognitive acts. Particularly that intuitions are not triggers for judgement making that are "fixed and stereotyped responses to environmental conditions that are unaffected by learning and experience" (Woodward 2016: 93). Rather intuition based judgement making is a form of cognition that is moulded by reflection. For instance, Hanno Sauer (2015: 161–163, 2012: 266–270) has shown that there is empirical evidence that moral intuitions are not merely modified by other people's judgements, but also by reflection that builds up considerations that speak in favour of or against particular intuitions.

Contesting Haidt's theory on empirical grounds like this is not meant to cast doubt on the very attempt to offer an empirical basis for moral judgement making. Rather the evidence merely shows that we do have reason to think that the use of reflective deliberative reasoning to make moral and political judgements is not as rare as Haidt claims and that Haidt's view of moral cognition is not as well supported by evidence as he claims. This does not mean that new empirical evidence could not possibly support Haidt's model of the human mind and political judgement making. It only shows that the reasons we had to adopt Haidt's theory are not as strong as they first seemed when we considered judgement change through reflection. It fails to explain a significant case set of reasonable disagreements. This is because it relies on a disjointed view of what goes on in people's minds when they make moral and political judgements. It relies on a disjointed model of moral cognition.

Given these empirical doubts, the more plausible view of moral cognition is that it involves cognitive acts on a spectrum from unreflective intuitive to reflective deliberative.[12] The more complex or greater number of inputs to the cognitive mechanism, by which I mean the host of desires, emotional dispositions, evidence and information that typically count as deliberative considerations, the more reflective and deliberative our judgement making is. The fewer or less complex the inputs, the more unreflective and intuitive our judgement making is. In fact Prinz (2016: 67–68) and, Kennett and Fine (2009: 88–91) have argued the empirical evidence actually shows an integrative model of moral and political judgement making is right.[13] On this view, the brain

[12] See also LaFollette and Woodruff (2015: 457–459) on a similar point.

[13] See Patterson et al. (2012), Young and Dungan (2012), and Schuler and Churchland (2011) on this as well.

is seen as having areas with different functions, with moral and political judgements being caused by a mix of emotional dispositions and deliberative reasoning. This supports the more modest theory of moral cognition proposed by Shaun Nichols (2004: 27–29, 62, Ch6.). The idea being that the most plausible view of the psychological foundations of moral and political judgement making lie somewhere between the complete rationalist view and complete intuitionist view like Haidt's. That rather, it involves people possessing a normative theory of moral and political beliefs that are entwined with moral emotions such that depending on how the emotions are regulated and sorted, judgements according to the normative theory will range from intuitive to reflective.

Given the two problems in Haidt's response to cases of reflective judgement making and the countervailing evidence in favour of a more modest model of moral and political judgement making we have good reason to reject Haidt's theory and move to what I call the Conceptual Family of theories in the next section. This move is not about ignoring psychological evidence or trying to explain reasonable disagreement with abstract entities. Rather it is to accept the general view of moral cognition I have mentioned and looking at facts about the role of concepts in the cognitive process of judgement making.

5 Conceptual Family

So far I have argued that we have good reason to reject the Imperfection and Historical-Psychological Family of theories for explaining reasonable disagreement. In this section I consider an alternative that attempts to avoid those problems: the Conceptual Family. This family involves theories which cite facts about the role of concepts in moral and political judgement making as difference makers. This strategy involves proposing a model of the cognitive process of moral and political judgement making that treats concepts as the explanatorily significant element. This strategy, allows theories to assume a general view of moral cognition where a plurality of psychological and historical facts sit at the foundation of judgement making. But, rather than trying to identify any particular psychological and historical facts, theories cite facts about the role of concepts as a particular information structure in moral cognition, and how this role is affected by a plurality of more foundational psychological and historical facts.

The facts that these theories cite fall into two types. One type of fact is differences in how people use concepts to form conflicting conceptions to explain reasonable

disagreement. Conceptions are general beliefs, typically principles that describe requirements, that people hold about justice or morality generally.[14] People then use these conceptions to make specific judgements about institutions and outcomes. I call theories that cite this type of fact Concept Use theories. Another type of fact is differences in how people possess concepts, *and* how they use them to form conflicting conceptions. I call theories that cite this type of fact Concept Possession and Use theories.

Given all that, the rest of this section proceeds as follows. In §5.1 I argue Concept Use Theories, whilst avoiding the objections that faced the Imperfection and Historical-Psychological Family of theories cannot explain reasonable disagreements that are deep disagreements and so is explanatorily weak. In §5.2 I argue that this leaves Concept Possession and Use theories as comparatively the best explanation of reasonable disagreement. This is because it can avoid the objections that faced theories that take up the Imperfection and Historical-Psychological views, *and* explain reasonable disagreements that are deep disagreements. Given that, it offers the most powerful explanation of reasonable disagreement. But I argue, the only extant theory that instantiates the type of theory suffers two serious problems. I argue, however, that rather than giving us reason to reject Concept Possession and Use theories, it motivates us to look for a better instantiation of it, which I take up in the next chapter.

5.1 Concept Use

As I have mentioned, Concept Use theories cite facts about differences in how people use concepts to form conceptions (ie. general beliefs about what justice or morality requires) to explain reasonable disagreement. One useful starting point to understand what this involves is what John Rawls (2005: 54–58) gestured towards with the "burdens of judgement" as the sources of reasonable disagreement.[15] For Rawls, there are six sources of reasonable disagreement:

1. The complexity and conflicting nature of empirical evidence.

2. Differences in the weight people afford deliberative considerations for making judgements.

[14] For this distinction between concepts and conceptions, see Rey (1985, 1983) for its early use in the philosophy of language and mind and Rawls (2005: 14, fn.15, 1999: 5–6, 9) for its use in political philosophy.

[15] See also Vallier (2019: 19–25) on Hayek's "scales of value" as a similar starting point.

3. Differences in how we use concepts because of their inherent vagueness in certain hard cases.

4. Differences in people's personal experiences affect how they weigh evidence and values.

5. Differences in the kinds of deliberative considerations relevant to certain cases makes it difficult to weigh them against each other.

6. Social institutions are limited in the values they can embody which makes it difficult to weigh and prioritise some values over others.

These sources illustrate the sort of facts Concept Use theories cite in their explanatory models. They all involve various ways reasonable people's weighing of deliberative considerations is affected in such a way that people form conflicting conceptions, and make conflicting political judgements. The only issue is that Rawls offers no account of how the sources interact in a cognitive process where concepts are used to make moral and political judgements and why they are the sources that explain reasonable disagreement. This is the crucial departure point for Concept Use theories. They offer an explanatory model that explicates how the burdens operate in a cognitive process to affect how people use concepts to form conceptions.

Andrew Mason (1993) and Christopher McMahon (2009) propose theories of this variety. They propose explanatory models that cite differences in people's psychological dispositions and personal experiences, as the facts that make the difference between reasonable disagreement and reasonable agreement. This involves describing how psychological dispositions and personal experiences affect how reasonable people weigh up deliberative considerations such that they form conflicting conceptions, and then make conflict political judgements according to them. Mason and McMahon theories then differ on the details of how to cash out the cognitive process of using concepts to form conceptions in their respective explanatory models.

The key notion in Mason's explanatory model is the idea of *essentially contestable concepts*. Mason (1993: 58) unpacks this as follows:

> Key political concepts such as 'democracy' and 'social justice' are essentially contested. They accredit a complex, valued achievement. Different elements in this achievement may be weighted differently by different contestants.

The idea here is that some concepts admit conflicting ways of forming conceptions because they involve considerations that are sufficiently complex to such that they

permit reasonable people to weigh them differently. For Mason (1993: 59), explaining *why* this happens involves an explanation of why moral and political concepts are actually *contested*. Mason's explanation for this, after significant revisions and qualifications to W.B. Gallie's original account, is that, "When there is some measure of freedom of thought and expression, political disagreement will emerge because different uses of a number of political concepts are reasonable and, under these circumstances, there will be a diversity of rational and non-rational causes of political belief formation." The idea here is that when we assume people are reasonable and not under some external coercion to believe what they believe, there will be different rational and non-rational factors which affect how they weigh the deliberative considerations provided by the content of their concepts to then form conflicting political beliefs (ie. conceptions).

But what are the "rational and non-rational" factors that affect how people form conceptions? These are, as Mason (1993: 15, 99–100) says, "the reasons that people have for making the judgements" or in other words their deliberative considerations, and the "psychological propensities or personal experiences" that affect how parties weigh the deliberative considerations provided by a concept. This means that even when reasonable people make no mistake in reasoning, they can form conflicting conceptions that entail conflicting judgements because of the differences in their psychological propensities and personal experiences that affect how they weigh the deliberative considerations associated with some concept. Given that these propensities and personal experiences cannot be erased, they give rise to conceptions that conflict to such an extent that they lead to judgements that conflict intractably.

In sum, Mason's explanatory model cashes out the cognitive process of using concepts to form conceptions with the idea that some moral and political concepts are essentially contestable. This means that what explains reasonable disagreement are differences in reasonable people's psychological propensities and personal experiences causing them to weigh the deliberative considerations provided by their moral and political concepts in different ways. When, for instance, the concept JUSTICE is contested in this way reasonable people form conflicting conceptions of justice which in turn entail intractably conflicting political judgements about what institutions and outcomes justice requires.

Another way to explain reasonable disagreement is to cash out the cognitive process of using concepts to form conceptions by focusing on how reasonable people learn how to use moral and political terms like "justice". This is what McMahon (2009: 3–4) does with his novel metaethical theory of "moral nominalism". McMa-

hon explains the basic idea of moral nominalism when he says:

> I have chosen this label because the possession of a moral concept is understood to consist in the mastery of the use of a moral term, which in turn is explained without invoking moral properties.

On this view, what is relevant for understanding the cognitive process that leads people to make moral and political judgements is how reasonable people learn how to use moral and political terms. For McMahon (2009: 55–58) learning how to use moral and political terms involves developing both an association with the set of features of a state of affairs that the moral and political terms apply to and a set of "extrapolative dispositions" for using the term. The features are a set of morally relevant deliberative considerations that a person learns to associate with a term's concept. The "extrapolative dispositions" are dispositions individuals have for weighing the deliberative considerations and forming conceptions or altering them when they encounter new states of affairs and have to use the term in question again. When reasonable people must use a moral or political term to make a judgement, their extrapolative dispositions will motivate them to extrapolate the use of a concept from what they have learned from their past personal experiences to the new case. The particular way they use their concepts to either confirm or alter their conceptions will depend largely on the extrapolative dispositions they develop and the particular personal experiences they have where those dispositions are activated or developed further.

On this model, reasonable people disagree because the precise nature of the people's personal experiences and the extrapolative dispositions they develop are what determine how they use their moral and political concepts to form conceptions. The conceptions reasonable people form by using their concepts will conflict because reasonable people have different personal experiences of learning how to use moral and political terms so that they develop different extrapolative dispositions for using the concepts associated with the terms. As McMahon (2009: 77) says of the dispositions:

> The extrapolative dispositions, that, according to moral nominalism, underlie the correct use of normative and evaluative terms will differ to a certain extent from person to person, even if the cognitive and motivational capacities of the people involved are functioning properly and they are employing a common set of terms. The dispositions may converge in a particular case, but there is no rational requirement that they do so.

Reasonable people will develop different dispositions because reasonable people will

differ in the personal experiences where they learn how to use moral and political terms. As McMahon (2009: 78) says:

> Because of their different experiences, they are likely, even when reasoning competently, to respond in different ways to the cases in the common set constructed by shared deliberation. It is partly because large modern polities contain people who have diverse personal experiences that reasonable disagreement has a prominent place in the life of such polities.

On McMahon's model then, reasonable people make conflicting moral and political judgements because they differ in their extrapolative dispositions and their personal experiences with regard to learning how to use moral and political terms. These differences in personal experiences and extrapolative dispositions affect how reasonable people use their concepts to form their conceptions of those moral and political concepts, which then leads to making conflicting moral and political judgements. This difference in how reasonable people use their concepts to form conceptions is then what makes the difference between reasonable disagreement and reasonable agreement. Since reasonable people cannot change or choose the circumstances in which they develop their extrapolative dispositions or, largely, the nature of experiences they have where they are forced to alter their conceptions, any conflict in judgements ends up an intractable conflict (McMahon 2009: 55–63, 78, 81). Learning how to use moral and political concepts involves experiences with parents and one's wider society that involve developing extrapolative dispositions. These early circumstances are largely outside a person's control and go on to determine the way they make their political judgements in the future.

In sum, like Mason's model, McMahon's model explains reasonable disagreement about justice by citing differences in reasonable people's psychological dispositions and personal experiences causing them to weigh the deliberative considerations provided by their concept of JUSTICE in different ways. On McMahon's model, these dispositions are the extrapolative dispositions that he proposes are developed when individuals learn how to use moral and political terms like justice. It is this learning process and the use of concepts it involves that causes reasonable people to form conflicting conceptions of justice and in turn make intractably conflicting political judgements about what institutions and outcomes justice requires.

To get a clearer idea of how Mason and McMahon's models explain reasonable disagreement about justice, consider the following case:

> *Abortion:* Barry and Nora are discussing the laws concerning abortion

> that their society ought to enact. Barry judges their society ought to enact laws that permit abortion in the first trimester because he believes respecting women's bodily autonomy is what justice requires. This is because respecting women's bodily autonomy outweighs the value of human life in a fetus. Nora, on the other hand, judges their society ought to outlaw abortion in the first trimester barring exceptional circumstances because she believes preserving the value of human life is what justice requires. This is because the instrumental and non-instrumental value of human life outweighs women's bodily autonomy.

Concept Use theories like Mason and McMahon's explain this case by citing facts about the cognitive process involved in reasonable people making their political judgements. They cite a difference in Barry and Nora's psychological dispositions and personal experiences that lead them to weigh their deliberative considerations differently and form conflicting conceptions of justice. Mason would cite differences in Barry and Nora's "psychological propensities and personal experience" with respect to abortion and pregnancies, whilst McMahon would cite their "extrapolative dispositions" and their learning experiences for the term "justice", "autonomy", and perhaps "murder". In both models, the point is that Barry and Nora's different personal experiences cause them to be disposed to weigh the value of life and bodily autonomy differently, and as a result form conflicting conceptions of justice. This causes them to make conflicting judgements about the institutions and outcomes related to abortion that justice requires.

This type of explanation covers cases of reflective and unreflective judgement making because in both the structure of the cognitive process is the same except for how much effort reasonable people put into considering the weights they give their deliberative considerations. The more they rehearse their deliberations and question how to weigh certain considerations, the more reflective their judgement making. The less they rehearse and question, the more unreflective and intuitive. The point here is that Concept Use theories can easily explain cases like *Abortion* because they cite facts about the cognitive process that causes parties to make conflicting judgements about what justice requires. This explanation works regardless of whether a particular act of judgement making involves reflection or not.

Taking stock for a moment, it is worth considering how Concept Use theories like Mason and McMahon's improve on the Imperfection and Historical-Psychological family of theories. The first improvement is that, Mason and McMahon's theory avoid all the problems that face the Imperfection Family. They manage to explain

reasonable disagreement without ruling it out and without assuming a normative standard that will itself be the subject of reasonable disagreement. This is because facts about how concepts are used to form conceptions are not the sort of facts that speak to any form of intellectual error.

The second improvement is that Mason and McMahon's theories can easily explain cases of reflective judgement making. This is because they assume a general theory of moral psychology of the sort argued for by Nichols (2004). This means they need not take any controversial empirical positions about the psychological foundations of moral and political judgement making. It merely has to cite what reasonable parties do with their concepts when they reflect on the judgements they have made and how this is affected by their psychological dispositions and personal experiences. To that extent, reflective judgement making simply involves reasonable people rehearsing and reconsidering how they weigh their deliberative considerations. If they differ in their dispositions that affect how they weigh certain deliberative considerations this will result in slightly different conceptions and hence different moral and political judgements. That people can reflect in this way, change their judgement and end up in disagreements is easily explained.

But despite all this, Concept Use theories face a major explanatory weakness. Not all cases of reasonable disagreement are like *Abortion*. Some are 'deep disagreements' and Concept Use theories cannot explain them. Cases of reasonable disagreement that are deep disagreements involve a conflict at a deeper level of thought, namely *about* the deliberative considerations that ought to be used in the cognitive process of moral and political judgement making. These are cases where reasonable people conflict about what the deliberative considerations with respect to some moral or political issue *are*, and not merely about how to weigh those considerations. Whilst Mason and McMahon's theories can explain the latter they cannot explain the former. Theorists have described this sort of disagreement as a conflict of, "worldviews", "perspectives" or "fundamental commitments".[16] There are generally two types of deep disagreement. The first type are disagreements like the following:

> *Nationalisation:* Bryan and Elizabeth are discussing the economic structure their society ought to have. Elizabeth judges their society ought to nationalise, at the very least, some key industries because a society's productive capacity being for the mutual benefit of all is what justice

[16] See Ranalli (2018b: 2–4, 2018a: 1–2), Hazlett (2014: 12–13), Pritchard (2018), Kappel (2018), Adams (1985) and Fogelin (1985) for an overview of deep disagreement in epistemology, and Gaus (Gaus 2018, 2017, 2016) and Muldoon (2016) in political philosophy.

> requires. Bryan judges their society ought not to nationalise any industries because protecting people's natural moral right to their body and private property is what justice requires. This is because he believes natural moral rights are what matter, benefiting everyone with society's productive capacity is irrelevant and has nothing to do with what justice requires. On the contrary, Elizabeth believes materially benefiting people is what matters, natural rights to private property are irrelevant and have nothing to do with what justice requires.

This is a case of a reasonable disagreement about justice that is a *Direct Deep Disagreement* because it involves reasonable people disagreeing about the considerations that ought to be used to deliberate about what justice requires. As such it is a disagreement that involves a conflict about the inputs to the cognitive process of deliberation that leads to people forming conceptions of justice and eventually judgements about what justice requires. Contemporary examples of this type of reasonable disagreement are those between those with distinct moral views on the political spectrum from libertarians, socialists, religious integralists and liberals. It is not disagreement within those groups but between them where each thinks the other is mistaken about what the relevant deliberative considerations are for deliberating about what justice requires.

Another type of deep disagreement that is subtly different to *Nationalisation* is the following:

> *Indirect Abortion:* Barry and Nora are discussing the abortion laws their society ought to enact. Nora judges their society ought to outlaw abortion because she believes, although it does restrict women's autonomy, fetuses are innocent persons which means killing them is murder and laws against murder are what justice requires. Barry, on the other hand, judges their society ought to enact laws that permit abortion in the first trimester because, whilst he agrees justice requires laws against murder, he believes fetuses in the first trimester are not people, so killing them is not murder and so permitting abortion to protect women's autonomy is what justice requires.

This is a case of *Indirect Deep Disagreement* because it involves reasonable people disagreeing *indirectly* about the considerations that ought to be used to deliberate about what justice requires. This captures the thought that Barry and Nora disagree about the considerations that ought to be used for deliberating about justice, by way of disagreeing about the content of a particular consideration and not about whether that

consideration ought to be used at all. For instance, both Barry and Nora agree that the 'no-killing-moral-persons' consideration ought to be used for deliberating about justice, but they disagree about its content. They disagree about who counts as a moral person. As such, it is still a disagreement that involves a conflict about the inputs to the cognitive process of deliberation that leads to people forming conceptions of justice and eventually judgements about what justice requires. Examples of this sort of disagreement go beyond abortion and moral personhood. Many contemporary disagreements like that over the use of female or woman, a religious or civil idea of marriage, and a structural or individualistic idea of coercion in law and morality are cases of *Indirect Deep Disagreement*. Many reasonable disagreements about justice hinge on a disagreement about those choices.

The point of deep disagreements like *Nationalisation* and *Indirect Abortion* is that they place Concept Use theories in a dilemma. On one horn, if they accept they are genuine cases of disagreement, then they have to concede that they do not have the explanatory resources to explain them. And, therefore that they have a serious explanatory weakness. This is because the disagreements involve a conflict about what the deliberative considerations ought to be for deliberating about some moral or political matter. This conflict is not a difference in *how* reasonable people perform this deliberation, but a difference in what they view as the appropriate inputs to that deliberation itself. Mason and McMahon's theories, however, do not have the resources to explain this. Their entire focus, in extending Rawls's burdens of judgement, was to explain reasonable disagreement by citing facts about how people weighed the deliberative considerations differently and so formed slightly different conceptions of justice. Since these conceptions involve general beliefs and principles about what justice requires, this would cause people to make different judgements about what justice requires. But, cases of deep disagreement outstrip the resources of that explanation. In cases of deep disagreement, it is perfectly possible that people weigh their deliberative considerations the same and yet have conflicting views about which considerations to weigh.

One the other horn, Concept Use theorists could deny that deep disagreements are actually genuine. They could say that they are verbal disagreements because they do not contain any conflict over the first-order issue that their judgements are about. Rather the disagreements involve a conflict about the deliberative process that ought to be used when making moral and political judgements. But, the parties to the disagreement both agree about what moral and political judgements are correct relative to each other's deliberative considerations and as such do not conflict over the truth

of each other's judgements. To that end, it does not matter that Concept Use theories do not have the explanatory resources to explain them. They are not genuine disagreements at all. What the parties in deep disagreements ought to do is resolve their verbal disagreement first and then come back to the table.

The problem with this response is that it is deeply counterintuitive. We have a strong intuition, or at least I do, that cases of deep disagreement are genuine disagreements about first-order moral and political issues. In *Nationalisation* and *Indirect Abortion* the parties explicitly deny their interlocutor's judgement about what justice requires. They conflict on the truth or correctness of these judgements. But, as stipulated, they conflict in this way in virtue of a disagreement about what the considerations ought to be used for deliberating about justice. This is because parties to the disagreements know that what considerations are used to deliberate about justice will directly affect the moral and political beliefs they form about the requirements of justice and how people act given those beliefs. Moreover, many of our most contentious disagreements in contemporary political life appear to be like this. Unlike paradigmatic verbal disagreements they seem to be worth having, and many of the participants in them on both sides seem to find them worth having, precisely because they are seen to matter for how coercive political power will eventually be justified or used. To that end, the Concept Use theories must concede that in denying that reasonable disagreements which are deep disagreements are genuine, they are committed to a deeply counterintuitive result about many cases of reasonable disagreement.

Where does this leave us? It leaves us, I submit, with good reason to reject Concept Use theories. Although they are an improvement on the Imperfection and Historical-Psychological family of theories, they face a dilemma between explanatory weakness, and counterintuitively explaining away deep disagreements as verbal disagreements. What can avoid this dilemma is, I submit, theories with a better explanatory model that can actually explain cases of deep reasonable disagreement. It is to this type of theory I turn to in the next section.

5.2 Concept Possession and Use

In the last section I mentioned that one way to avoid the dilemma that Concept Use theories face as a result of deep disagreements is to find a better way to explain reasonable disagreement by citing facts about the role of concepts in people's moral and political making. This is where Concept Possession and Use theories come into play. They propose an explanatory model that cites how the possession of a concept in addition to its use affects how people form conceptions and as a result make moral and

political judgements.

The core idea is that this way of cashing out the role of concepts in the cognitive process of moral and political judgement making can explain cases of reasonable disagreement that are deep disagreements. This is because facts about how people possess concepts will be facts about the deliberative considerations a concept provides when forming conceptions. Given that, deep disagreements involve a conflict over what the relevant deliberative considerations are when making moral and political judgements about some particular issue, facts about how people possess concepts would explain this conflict.

Ronald Dworkin's theory of interpretive concepts offers an explanation of this sort by cashing out concept possession and use with the idea of possessing and using interpretive concepts. The core idea being to distinguish the type of concept involved in reasonable disagreements as a way to individuate the facts about concept possession and use that make the difference between reasonable disagreement about reasonable agreement. As such, Dworkin's (2011: 159–160) theory starts with making a distinction between "criterial", "interpretive" and "natural-kind" concepts on the basis that it provides the best way to make sense of how reasonable people can have conflicting views of what the deliberative considerations ought to be when forming conceptions.[17] The best way to understand the distinction and its explanatory role is to contrast what Dworkin thinks it means to use criterial concepts as opposed to interpretive concepts. Dworkin (2011: 160) says of criterial concepts:

> ...criterial and natural-kind concepts do have something important in common. People do not share a concept of either kind unless they would accept a decisive test – a kind of decision procedure – for finally deciding when to apply the concept (except in cases they agree are marginal). Genuine disagreement about application is ruled out once all pertinent facts are agreed upon.

The idea here is that criterial concepts are concepts where their use has a precise criteria and the acceptance of the same criteria for use determines whether reasonable people share the concept. On this view, states of affairs or objects fall into the extension of the respective criterial concept if they meet the criteria for that concept. Facts about those states of affairs and objects will determine whether they meet the criteria. This means that a disagreement where the conflicting judgements are made using criterial concepts is explained by the fact that either, the disagreeing parties do not share the

[17] Nothing in how Dworkin's theory explains reasonable disagreement hinges on natural kind concepts and so I will not discuss them any further.

same criteria in which case it is a verbal disagreement between two different concepts, or the disagreeing parties do not agree on facts about the world that would determine whether something meets the criteria they both share in which case at least one of the disagreeing parties would be making a straightforward intellectual error. Given that, reasonable disagreements cannot be explained by facts about how we possess and use criterial concepts. Such an explanation would succumb to all the problems that plagued the Imperfection family of theories, or implausibly rule out all reasonable disagreements as verbal disagreements.

This is why Dworkin proposes that moral and political concepts are not criterial. Moral and political concepts – the concepts available to individuals when making political judgements about justice – are interpretive. Dworkin (2011: 6) describes interpretive concepts generally, when he says:

> We must therefore recognize that we share some of our concepts, including the political concepts, in a different way: they function for us as interpretive concepts. We share them because we share social practices and experiences in which these concepts figure. We take the concepts to describe values, but we disagree, sometimes to a marked degree, about what these values are and how they should be expressed. We disagree because we interpret the practices we share rather differently: we hold somewhat different theories about which values best justify what we accept as central or paradigm features of that practice. That structure makes our conceptual disagreements about liberty, equality, and the rest genuine.

There are two features of interpretive concepts that need to be teased out from that passage. The first is that for Dworkin, moral political concepts as interpretive concepts describe values. These values involve an interpretation of what is of value and ought to be furthered in a particular social practice of using that concept. To that end, the content of interpretive concepts involve values that are the result of an interpretation of the function of the concept itself.

The second feature of interpretive concepts is that if their content are values identified by interpreting the function of using that concept in a social practice, then to use an interpretive concept is to engage in conceptual interpretation. It is to identify what is of value or disvalue by interpreting the purpose of identifying what is of value or disvalue in the social practice that we are engaged. This interpreting then establishes the content of the concept and the values that people use to form conceptions and making judgements. But, Dworkin (2011: 162–163) does warn that there is a non-

foundationalist circularity to his theory. In specific reference to the interpretation of moral concepts he says:

> But can interpretive arguments about justice escape a narrow circularity?...There is no circularity in interpreting a statute by supposing it to serve the value of equality. But moral concepts themselves designate values. How can someone identify the value latent in the practices of justice without appealing, unhelpfully, to the concept of justice itself?...We defend a conception of justice by placing the practices and paradigms of that concept in a larger network of other values that sustains our conception. We can in principle continue this expansion of our argument, exploring other values until, as I said, the argument meets itself. The circularity, if any, is global across the whole domain of value.

On this view conceptual interpretation involves the possession of other interpretive concepts. This is what Dworkin (2011: 154) means when he says that conceptual interpretation is "pervasively holistic". The use of interpretive concepts relies on the interpretation of a network of concepts we possess. As Dworkin (2011: 154) says:

> Interpretation is pervasively holistic. An interpretation weaves together hosts of values and assumptions of very different kinds, drawn from very different kinds of judgment or experience, and the network of values that figure in an interpretive case accepts no hierarchy of dominance and subordination. The network faces the challenge of conviction as a whole; if any one strand is changed, the result may be locally seismic.

The result of this is that the use of an interpretive concept involves relying on the network of other interpretive concepts.[18] This means what is identified as valuable or disvaluable in a shared social practice will rely on people using other interpretive concepts and so any single act of interpretation is guided by how it fits with the entire web of interpretations.[19]

At this point we might wonder how reasonable people can be said to possess and use the same concept if they differ in their interpretation of what they are doing with the concept? To solve this Dworkin proposes that in the case of interpretive concepts what it means to have "an understanding that their correct application is fixed by the

[18] The point here is that the circularity is justified by Dworkin's stronger thesis about the independence and holistic unity of the true interpretations of all our values. See Winter (2016) and Knight (2006) for discussion of this view.

[19] See Plunkett and Sundell (2013b: 251–252) for this broad understanding of Dworkin's theory.

best interpretation of the practices in which they figure" is that all agree on certain paradigm uses of the concept. As Dworkin (2011: 160–161) explains:

> People participate in social practices in which they treat certain concepts as identifying a value or disvalue but disagree about how that value should be characterized or identified. The concept of justice and other moral concepts work in that way for us....We do not agree about what makes an act just or unjust, right or wrong, an invasion of liberty or an act of tactlessness. But we agree sufficiently about what we take to be paradigm instances of the concept, and paradigm cases of appropriate reactions to those instances, to permit us to argue, in a way intelligible to others who share the concept with us, that a particular characterization of the value or disvalue best justifies these shared paradigms.

The point here is that sufficient agreement about paradigm uses of a concept, which for Dworkin means particular judgements, will ensure that everyone is using the same interpretive concept. This will be the case even if people interpret the content of the concept so differently they hold vastly different beliefs about what the appropriate values are when engaging in the share social practice of using that concept.

Putting all this together we get the following explanation of reasonable disagreement. The facts that make the difference between reasonable disagreement and reasonable agreement are facts about conceptual interpretation. This is because the concepts that feature in reasonable disagreements are moral and political concepts which in turn are interpretive concepts. This means for Dworkin's theory the difference makers are facts about how reasonable people possess different interpretations of the moral or political concept they use, to make moral or political judgements. The fact that they differ in their interpretation of the concept means they will make conflicting moral and political judgements. But, what are these facts about differences in the conceptual interpretation that cause people to interpret a moral or political concept differently? The relevant facts on Dworkin's view are an individual's experiences and the innate dispositions that affect their conceptual interpretations across their network of interpretive concepts. As Dworkin (2011: 150-151) says:

> Disagreement is patent, but its source almost always is obscure, buried in a large variety of unarticulated assumptions about law or art or literature or history that rarely surface and that can be explained only as the upshot of some combination of inherent taste, training, acculturation, allegiance, and habit.

We can for the sake of simplicity group "inherent taste" and "habit" under the label of psychological dispositions, and "training, acculturation, allegiance" under the label of personal experiences. This means that differences in conceptual interpretation involve differences in people's psychological dispositions and personal experiences relating to the interpretation of a concept. In the case of a concept like JUSTICE, differences in people's psychological dispositions and personal experiences affect how they interpret the purpose and aim of the social practice of using the concept of JUSTICE. For Dworkin (2011: 162) this social practice is one where people theorise and make judgements about what society's institutions ought to be like and how individuals ought to be related to one another under those institutions. But, recall that interpreting the purpose and aim of a social practice that involves using the concept JUSTICE involves relying on one's network of interpretive concepts and the values they describe. This means that reasonable disagreement about justice is produced because reasonable people's different psychological dispositions and personal experiences cause them to rely on the network of their interpretive concepts in slightly different ways. This difference will then support forming conflicting conceptions of justice and making conflicting judgements about the institutions and outcomes that justice requires.

This means that ordinary cases of reasonable disagreement like *Abortion* are explained by differences in reasonable people's psychological dispositions and personal experiences causing them to interpret the concept of JUSTICE such that they weigh a shared set of values in different ways. In these cases reasonable people share the interpretation of their network of interpretive concepts to a sufficient degree that they identify the same values as the best interpretation of the aim and purpose of the social practice of using justice as a concept. But, their psychological dispositions and personal experiences cause them to weigh these values differently such that they form conflicting conceptions of justice and make conflicting political judgements according to them.

In the case of reasonable disagreements that are deep disagreements, the explanation is that reasonable people differ in the interpretation of their network of interpretive concepts such that they identify different values as the best interpretation of the aim and purpose of the social practice of using JUSTICE as a concept to form conceptions of justice. This is what explains how a reasonable disagreement can turn on a conflict about what the appropriate inputs to deliberation are for forming conceptions of justice and making judgements about what justice requires. In the case of a *Direct Deep Disagreement* like *Nationalisation*, the disagreement is explained

by differences in reasonable people's dispositions and personal experiences causing the parties to identify different values as relevant for the best interpretation of the concept of JUSTICE. But, in the case of a *Indirect Deep Disagreement* like *Indirect Abortion*, the disagreement is explained by differences in reasonable people's dispositions and personal experiences causing the parties to interpret the concept of MORAL PERSONHOOD differently and therefore, in virtue of that, interpret JUSTICE slightly differently.

In sum then, on Dworkin's theory reasonable disagreement is explained by differences in reasonable people's psychological dispositions and personal experiences affecting their interpretations of the moral or political concept they are using. The differences result in either weighing shared values differently or identifying different values as the best interpretation of the aim and purpose of making judgements using the concept in question. This way of cashing out the cognitive process of how reasonable people use moral and political concepts to make moral and political judgements improves on Concept Use Theories whilst maintaining the advantages over the Imperfection and Historical-Psychological families. It improves on Concept Use theories because it allows for an explanation of reasonable disagreements that are deep disagreements as genuine disagreement. And, it does this, whilst still being able to explain reflective judgement making unlike theories in the Historical-Psychological family, without any notion of intellectual error like the Imperfection family.

But the improvement over Concept Use theories comes at a cost. Specifically, Dworkin's theory of interpretive concepts faces two problems, the Virtue of Reasonableness Problem, and the Regression of Interpretation Problem. The Virtue of Reasonableness Problem is that since on Dworkin's theory all moral and political concepts are interpretive concepts, reasonableness itself will end up being an interpretive matter and therefore, whether Dworkin's theory explains reasonable disagreement will itself be the subject of reasonable disagreement. The Regression of Interpretation Problem is that, given the nature of conceptual interpretation, when reasonable people in deep disagreements try to communicate why they make their judgements, Dworkin's theory entails that they will descend into a regression of interpretation. Both problems show that we have good reason to reject Dworkin's theory.

The Virtue of Reasonableness Problem arises because on Dworkin's theory, the notion of reasonableness is, or at least partly, a moral concept. Recall from Chapter 1, that part of the idea of people being reasonable was that they had a minimal capacity for sincerely making judgements that they think others can agree to. For Dworkin (2011: 102–103) this is captured by the idea of "moral responsibility". As Dworkin

says:

> In this chapter we consider moral responsibility as a virtue. We begin with one aspect of that virtue. Morally responsible people act in a principled rather than an unprincipled way; they act out of rather than in spite of their convictions. (Dworkin 2011: 103)
>
> Count the ways in which someone might fail to act out of the principles he professes. The most obvious is crude insincerity. The leader who takes his country to war pretending to follow principles that in fact have no grip on him, principles that he has no intention of following when it is inconvenient for him to do so, is crudely insincere. (Dworkin 2011: 104)

On this view, "moral responsibility" is what captures at least part of the idea of people being reasonable. They have a certain moral character where they make sincere judgements and not ad-hoc or unprincipled ones. This fits with the way Dworkin describes moral responsibility as a virtue and by extension a behaviour or disposition to behave in a way that is valued. The value of sincere moral and political judgement making is furthered by having coherent judgements. If this is the case it seems clear that reasonableness, or at least part of it, is an interpretive concept. This is because Dworkin proposes that all moral and political concepts are interpretive concepts.

But this is a problem. If reasonableness is an interpretive concept then it will be a matter of interpretation whether there is any reasonable disagreement. If this is the case, then whether Dworkin's theory can actually explain reasonable disagreement will be a matter of interpretation. But, this would make the explanation viciously circular. This is because whether the explanation works would be a matter of reasonable disagreement. Recall, that to count as an explanation is to describe the facts that make the difference between reasonable and unreasonable disagreement. If that difference is itself interpretive, then Dworkin's theory will itself be the subject of reasonable disagreement and working out whether it works would require conceptual interpretation. This will in turn mean, whether a political judgement is reasonable will be decided by an individual's interpretation. But, without a shared interpretation, it will mean there is no genuine case of reasonable or unreasonable disagreement. In effect, a part of the explanandum – a disagreement's reasonableness – would be part of the supposed explanans – Dworkin's theory – and so there is no non-circular way of identifying a genuine case of reasonable disagreement.

Aside from the Virtue of Reasonableness Problem, there is another even more worrying problem for Dworkin's theory, the Regression of Interpretation Problem. This problem arises when reasonable people try to communicate their reasons for why

they make their judgements they must interpret their concept and communicate this act of interpretation to others. After all according to Dworkin's theory the main reason people make the moral and political judgements they do is because they interpret a concept a particular way. But, since interpretation is as Dworkin (2011: 154) says "pervasively holistic", communicating the interpretation of any one moral concept will mean having to communicate the interpretation of our entire network of interpretive concepts. But, of course at some point this will run out and they will have to communicate how they interpret the idea of "interpretation" itself. This is because, as Dworkin (2011: 131) says, "Interpretation is therefore interpretive, just as morality is moral, all the way down". If interpretation itself is an interpretive concept, then the only way to communicate it in a non-circular way is to interpret what one does when one interprets "conceptual interpretation". But that concept, in virtue of being what it is connects to all of one's other interpretive concepts. Since Dworkin does not place any principled constraints on when interpretation ends, there seems to be a clear danger that Dworkin's theory will entail a regression of interpretation.

But this is clearly not what happens in cases of reasonable disagreement. It seems clear that the disagreeing parties do manage to communicate why they make their respective judgements and have their interlocutors understand this reason. In fact this was what deep disagreements are predicated on. Disagreeing parties do understand their interlocutor's reasons because they reject them as relevant deliberative considerations. It is then implausible that parties descend into a regression of interpretation that leads to them never being able to intelligibly communicate what they are saying and why they are saying it.

The right response to all this is, I think, to concede that although we have good reason to reject Dworkin's theory, that reason does not speak against Concept Possession and Use theories per se. This is because the underlying issue in both the Virtue of Reasonableness and Regression of Interpretation Problems, is that in an effort to cash out facts about how concept possession and use cause reasonable disagreement, Dworkin takes the extreme step of proposing an entirely new type of concept, interpretive concepts. It is the wide scope of these concepts – covering all moral and political concepts – that causes the Virtue of Reasonableness Problem.[20] And, it is the way 'conceptual interpretation' makes the content of all interpretative concepts interconnected that gives rise to the Regression of Interpretation Problem. But, a Concept Possession and Use theory need not rely on such an idiosyncratic theory of

[20] See Plunkett and Sundell (2013b: 253–255) on this problem of Dworkin's theory giving a very wide scope to what can qualify as an interpretive concept.

concepts. There is nothing inherent to the idea of explaining reasonable disagreement by citing facts about how concepts are possessed and used that weds us to proposing an entirely new and empirically unsupported type of concept.

Dworkin's theory also shows that Concept Possession and Use theories are the best type of theory for explaining reasonable disagreement. After all, Dworkin's theory does explain deep disagreements without falling prey to the problems faced by the Imperfection and Historical-Psychological theories. As such it is the most explanatorily powerful theory we have looked at so far. This shows that citing facts about how reasonable people possess and use concepts, is the most explanatorily powerful way of explaining reasonable disagreement.

All this means the promise of Concept Possession and Use theories can be salvaged. But, what we require is a theory that preserves the improvements of Dworkin's theory and does not succumb to the problems that it faces. Salvaging Concept Possession and Use theories is what I turn to in the next chapter. I will argue for a novel theory that makes use of recent developments in the philosophy of language on metalinguistic negotiations. This follows a similar attempt by Plunkett and Sundell (2013b) to make use of the idea of metalinguistic negotiations to show how Dworkin's theory of interpretive concepts is not required to explain legal disagreements. In the next chapter I attempt the same task with with reasonable disagreement.

6 Conclusion

The point of this chapter was to argue in favour of Concept Possession and Use as a type of theory for explaining reasonable disagreement about justice. I did this by first arguing that theories in the Imperfection and Historical-Psychological Families face a range of problems that show they are inadequate explanations. I argued the Imperfection Family of theories either cite facts that do not make a difference between disagreement and agreement, or fail to justify the normative standard that would pick out the fact that does make a difference. I then argued the Historical-Psychological Family of theories cannot explain cases of disagreement that arise when people change their judgement through reflection.

This then motivated the Conceptual Family of theories which held the promise of avoiding the problems that faced the Imperfection and Historical-Psychological Families, by citing facts about the role of concepts in the cognitive process that leads reasonable people to make their moral and political judgements. I argued that although Concept Use theories make good on this promise they face a dilemma when faced

with cases of reasonable disagreement that are deep disagreements. They either have to acknowledge them as genuine disagreements in which case they have to admit they do not have the resources to explain them, or they have to counterintuitively deny they are genuine.

I then argued that Concept Possession and Use theories avoid the dilemma by accepting deep disagreements are genuine and citing differences in how people possess a concept itself as well as how they use it as difference making facts. Whilst this yielded the most powerful explanation, I argued the only extant theory that instantiates it – Dworkin's theory of interpretive concepts – faces two serious problems. I argued that it makes the reasonableness of a disagreement the subject of reasonable disagreement itself, and entails that any attempt by parties to communicate why they make their judgements results in a regression of interpretation.

But I argued this does not condemn the Concept Possession and Use theories itself but only Dworkin's specific theory. This is because the problems Dworkin's theory faces stem from the theory's commitment to interpretive concepts as a completely novel and unique type of concept. But, nothing in the idea of Concept Possession and Use commits us to proposing such a concept. Therefore, Concept Possession and Use theories could be salvaged if we could find a theory that preserves the improvements of Dworkin's theory and does not succumb to the problems that it faces. This is what I turn to in the next chapter.

Chapter 3

Diverse Packages Theory

1 Introduction

In the previous chapter I argued for Concept Use and Possession as the best type of theory within the Conceptual Family of theories for explaining reasonable disagreement about justice. Theories of this type explain reasonable disagreement about justice by citing facts about the possession and use of concepts in moral and political judgement making. I showed theories of this type improve on Concept Use theories, the Imperfection family of theories and the Historical-Psychological family of theories.

But, all was not rosy for Concept Use and Possession theories. I argued that although they can explain cases of deep disagreement, unlike Concept Use theories, the way they do it leaves them open to the Virtue of Reasonableness Problem and the Regression of Interpretation Problem. Recall, the Virtue of Reasonableness Problem is the idea that Concept Use and Possession theories make the reasonableness of a given disagreement a matter of reasonable disagreement itself. The Regression of Interpretation Problem on the other hand is the idea that Concept Use and Possession theories entail that any attempt by reasonable people to communicate why they make their judgements will descend into a regression of interpretation.

But I concluded that all was not lost because Concept Possession and Use theories could be salvaged. This was because the problems it faces are not inherent to Concept Possession and Use theories, but unique to the only extant theory that instantiates it: Dworkin's theory of interpretive concepts. I argued that the way Dworkin's theory takes the extreme step of proposing interpretive concepts as a completely novel and unique type of concept is the source of the problems. But, if a theory could be developed that did not rely on such a theory of concepts the Concept Possession and Use

version could be salvaged.

In this chapter I attempt that salvage project by explicating and arguing for Diverse Packages Theory as a novel explanation of reasonable disagreement. The core idea of the theory being an explanatory model with two distinctive features. The first feature is that by making use of innovations in the philosophy of language, it can read cases of reasonable disagreement as either canonical disputes or as metalinguistic negotiations. The second feature is that it can explain why reasonable disagreements, read in one of those two ways, occur by describing how reasonable people possess and use 'diverse concept-conception packages'. I argue that this theory can explain cases of deep disagreement, and so retains all the explanatory power of Dworkin's theory, and can also avoid the Virtue of Reasonableness problem and the Regression of Interpretation Problem. Given that, it is the best explanation of reasonable disagreement.

The chapter proceeds as follows. In §2 I motivate and describe the two moving parts of Diverse Packages Theory. In §3 I apply the theory's explanatory model to the cases of reasonable disagreement that motivate it and show it explains them. In §4 I describe the advantages of Diverse Packages Theory over its competitors. In §5 I consider and respond to three potential objections against the theory.

2 Diverse Packages Theory

The aim of Diverse Packages Theory is to offer the best explanation of reasonable disagreement. Recall that this means explaining the following:

Reasonable Disagreement about Justice: A state of affairs of intractably conflicting judgements about the institutions and outcomes justice requires, made by at least two parties who both have, a minimal capacity for rationality and a minimal capacity for sincerely making judgements that they think others can agree to.

Like the other theories I have looked at so far, Diverse Packages Theory targets that explanandum by aiming to explain reasonable disagreements in general, no matter their topic. To that end, the core idea of Diverse Packages Theory is to propose an explanatory model that explains why reasonable disagreements occur by reading them in one of two ways. It can explain them, depending on the cases at hand either by reading them as canonical disputes, or by reading them as metalinguistic negotiations. The idea being that this allows the model to parse any case of reasonable disagreement, whether they are ordinary disagreements or deep disagreements, as a genuine disagreement. In short, it vindicates the intuition that these cases involve at least two people

who each hold mental content of some kind (eg. beliefs, plan, judgement) that conflict in such a way that they both cannot be true. The theory then explains why these disagreements occur by citing differences in how reasonable people possess and use 'concept-conception packages' to make their moral and political judgements. To get an idea of what that means and what sort of explanatory model it involves, in what follows I lay out its two moving parts.

2.1 Canonical Disputes and Metalinguistic Negotiations

The first moving part is the idea that reasonable disagreements are not all genuine in the same way. That is to say, by making use of Plunkett and Sundell's (2013a,b) innovative analyses of normative and evaluative disagreements, reasonable disagreements can sometimes be canonical disputes and sometimes they can be a particular form of non-canonical dispute, a metalinguistic negotiation.

When disagreements are canonical disputes they are genuine in virtue of a conflict in what speakers literally express. For any given case of disagreement, reading it as a canonical dispute then involves two suppositions. The first supposition is that it hinges on a conflict in the mental content that speakers literally express, or would literally express. The second supposition is that as result of the first supposition the speakers must mean the same thing by their words because otherwise they would not conflict in what they literally express. They would be expressing mental contents with different truth-conditions. To see the motivation for reading reasonable disagreements in this way, consider the following case from Chapter 2:

> *Abortion:* Barry and Nora are discussing the laws concerning abortion that their society ought to enact. Barry judges their society ought to enact laws that permit abortion in the first trimester because he believes respecting women's bodily autonomy is what justice requires. This is because respecting women's bodily autonomy outweighs the value of human life in a fetus. Nora, on the other hand, judges their society ought to outlaw abortion in the first trimester barring exceptional circumstances because she believes preserving the value of human life is what justice requires. This is because the instrumental and non-instrumental value of human life outweighs women's bodily autonomy.

This is a familiar sort of case that haunts much of our political life. The best way to explain why a case like *Abortion* is a genuine disagreement is to read it as a canonical dispute. This involves first positing that it involves a conflict in the mental content

the interlocutors literally express, and second, that as a result the interlocutors mean the same thing by their words. In the case of *Abortion*, there is clear evidence for making both suppositions. On the first it is clear that what Barry and Nora literally express, or more accurately what they *would* literally express, are conflicting beliefs about the abortion laws justice requires. In short, they express beliefs that cannot both be true. On the second, assumption, if Barry and Nora do literally express conflicting beliefs, they must mean the same thing by "justice". Otherwise their beliefs would not actually be in conflict because they would have different truth-conditions for beliefs about the abortion laws justice requires. If Barry and Nora mean different things by "justice" then the way they use the term could be perfectly compatible. This would mean that *Abortion* was not a disagreement at all. But given the evidence, it is clear that Barry and Nora express conflicting beliefs and therefore they must mean the same thing by their words.

But Diverse Packages Theory recognises that not all reasonable disagreements are like *Abortion*. By which I mean they are not all cases of disagreement that hinge on the beliefs that reasonable people would literally express. Rather, as we saw in Chapter 2, there are cases of deep disagreement. These are cases where reasonable people disagree about the deliberative considerations that ought to be used to make moral and political judgements. To get a sense of what this means consider again the following simple case of reasonable disagreement which is a deep disagreement we first saw in Chapter 2:

> *Nationalisation:* Bryan and Elizabeth are discussing the economic structure their society ought to have. Elizabeth judges the nationalisation, at the very least, of some key industries is just because a society's productive capacity being for the mutual benefit of all is what justice requires. Bryan judges the nationalisation of any industries is unjust because protecting people's natural moral right to their body and private property is what justice requires. This is because he believes natural moral rights are what matter, benefiting everyone with society's productive capacity is irrelevant and has nothing to do with what justice requires. On the contrary, Elizabeth believes materially benefiting people is what matters, natural rights to private property are irrelevant and have nothing to do with what justice requires.

Bryan and Elizabeth's disagreement does not seem to hinge on what they would literally express, namely their conflicting beliefs about the economic structure justice requires. Rather it hinges on what they think are the relevant deliberative consid-

erations for deliberating about justice. Bryan thinks the protection of natural moral rights is the relevant consideration (and nothing else), and Elizabeth thinks materially benefiting people is the relevant consideration (and nothing else).

This rings true not merely of *Nationalisation*, but also of many contemporary political disagreements. For instance consider disagreements over whether the United Kingdom ought to leave or remain in the European Union, or whether transwomen are women, or whether private healthcare insurance markets constitute freedom. All these disagreements seem to hinge not on the beliefs or judgements reasonable people literally express, but on something like their entire view of how reasonable people ought to go about deliberating when making moral and political judgements about those issues.

But all this presents a problem. If these cases of deep disagreement do not hinge on a conflict over what is literally expressed, then an explanatory model that only treats reasonable disagreement as a canonical dispute is compelled to rule such cases as merely verbal disagreements. They would not qualify as genuine disagreements. This is because canonical disputes are disagreements that are genuine in virtue of involving a conflict over what is literally expressed. But, this contradicts the strong intuition we have that cases of deep disagreement like *Nationalisation* and other contemporary political debates, are genuine. They are not disagreements about labelling parts of the world in different ways. They are disagreements where the reasonable people express views about what their social world ought to be like that cannot both be true.

Of course one way to go would be to declare that despite our intuition these deep disagreements are not in fact genuine disagreements. If they do not hinge on what is literally expressed then ipso facto they cannot possibly involve a conflict in mental contents. But, this is plainly unjustified. We do not have any reason, independent of any particular explanation of reasonable disagreement, that defeats our intuition that deep disagreements are genuine.

Instead of this, I propose making use of recent work in the philosophy of language on "non-canonical disputes". The idea being that one way to vindicate the intuition that cases of deep disagreement are genuine is, I submit, to follow what Plunkett and Sundell (2013a: 11–13, 2013b: 247–248) argue about normative and evaluative disagreements in general, and go beyond treating all reasonable disagreement as canonical disputes. We need to treat some cases of reasonable disagreement as non-canonical disputes. This means recognising that some cases of reasonable disagreement are genuine not in virtue of a conflict in the mental content that speakers *literally* express. Rather they are genuine in virtue of a conflict in the mental content that speakers *pragmati-*

cally express. This means that cases of deep disagreement like *Nationalisation*, do not involve reasonable people literally expressing some mental content. Rather they are pragmatically expressing some conflicting mental content that is nevertheless related to the subject matter of their discussion.

But this raises the question of what precisely is being pragmatically expressed in such disagreements? It certainly cannot be reasonable people's beliefs about the institutions and outcomes that justice requires. After all, those are literally expressed. For this, I propose we use what Plunkett and Sundell say about a special type of non-canonical dispute: a "metalinguistic negotiation". A metalinguistic negotiation is, as Plunkett and Sundell (2015: 837–851; 2013a: 13–18, 2013b: 256–266) have argued, a disagreement that is *non-canonical* in virtue of consisting of a conflict about what a word ought to mean. This means the content that is pragmatically expressed is a 'metalinguistic belief' about what the meaning of a word ought to be.[1] This also means that disagreeing parties mean different things by, at least, one of their words. This is because it is the use of words with different meanings that pragmatically expresses the belief about what the word ought to mean.

To get a better sense of what all that means, consider the following case:

> *Spicy Soup:* Oscar and Callie are cooking soup for a party and are in a heated debate about its spiciness. They both taste the soup and, Oscar with the taste palates of the party guests in mind judges that the soup is spicy. But Callie with her long experience tasting chillies in mind disagrees and judges the soup not spicy at all.[2]

The point of *Spicy Soup* is that we have a strong intuition that it is a genuine disagreement. But, if it is read as a canonical dispute it would not be. This is because it does not seem to hinge on what Oscar and Callie believe about the soup, but rather what they think the threshold for spiciness should be. To solve this we can read it as a metalinguistic negotiation. This means supposing two things. The first supposition is that Oscar and Callie's disagreement involves a conflict in the mental content that speakers *pragmatically* express, namely their metalinguistic beliefs about what the meaning of "spicy" ought to be. The second supposition is that, as a result of the first supposition, they must mean different things by "spicy". It is, after all, the use of "spicy" with different meanings that pragmatically, rather than literally, expresses their metalinguistic belief about what it ought to mean.

[1] See also Chalmers (2011: 522-523) on these implicit metalinguistic beliefs.

[2] This is a slightly modified case that Plunkett and Sundell (2013a: 14–15) use to illustrate the power of reading disagreements as metalinguistic negotiation.

Why should we suppose these things? With the first supposition, I agree with Plunkett and Sundell (2013a: 19–20) that one piece of evidence for this is whether disagreeing parties are likely to carry on disagreeing even when they both agree about what their terms currently mean in the community at large or according to some authoritative third-party. This clearly seems to be the case in *Spicy Soup*. Oscar and Callie seem disposed to carry on their disagreement even if they could agree on what "spicy" means to other third-parties or how it is defined in a recipe book. This is because many normative consequences could flow from whether it is resolved on Oscar or Callie's behalf. Amongst other things it could affect whether the soup ought to be served, whether it ought to be praised more or less by the party guests, or even whether Oscar and Callie ought to cook at all.

With the second assumption, I agree with Plunkett and Sundell (2013a: 15) that one piece of evidence for this is whether disagreeing parties are disposed to use their terms in different ways, or, as Plunkett (2015: 847) says whether "speakers are disposed to systematically use a term in divergent ways in the same (non-defective) conditions". More simply, one piece of evidence for the second supposition is that the meaning of our words depends on our patterns of using that word. To that extent Oscar and Callie do seem disposed to systematically use "spicy" in divergent ways in the process of cooking the soup. By this I mean they are not disposed to suddenly use the term in the same way to refer to the same things as "spicy". To that end, it seems plausible to conclude they mean different things by "spicy". All in all, given the first supposition, it seems that Oscar and Callie clear are disposed to use "spicy" in systematically divergent ways. Therefore, if word use is a guide to its meaning, it seems plausible to conclude that they mean different things by "spicy".

Now, with all that in mind, I propose extending the strategy of vindicating the intuition that a disagreement is genuine through the idea of metalinguistic negotiation to, and only to, those cases of reasonable disagreement that are deep disagreements.[3] This means only reading as metalinguistic negotiations those reasonable disagreements that hinge on a conflict about the considerations relevant for deliberating about what justice requires (like in *Nationalisation*), rather than for example a conflict about whether justice requires Tax A, as metalinguistic negotiations.

Before moving to more concrete examples, it is work getting a bit clearer on how this startegy is supposed to work. The idea is that the conflict that reasonable dis-

[3]Note, Plunkett and Sundell (2013a: 7, 18–25, 2013b: 265–266) believe a metalinguistic analysis generally (not merely metalinguistic negotiation) can be plausibly extended to all kinds of normative, and evaluative disagreements. But, they do not touch upon how it can be used in the context of reasonable disagreements and particular deep disagreements.

agreements that are deep disagreements hinge on – a conflict about the considerations relevant for deliberating about what justice requires – is itself the result of some first-order reasonable disagreement unrelated to the question of what justice requires (or morality generally). These first-order reasonable disagreements will of course be canonical disputes that hinge on various metaphysical, epistemic, axiological, logical, and scientific topics like the existence of certain properties, the nature of certain beliefs and knowledge, the value of certain properties or states of affairs, or even perhaps the correctness of certain logical systems. When someone deliberates about what justice requires using some consideration that features in these reasonable disagreements on the various metaphysical, epistemic, axiological, logical, and scientific topics they end up having, I submit, a metalinguistic negotiation. This means explaining the essential difference involved in a conflict about the considerations relevant for deliberating about what justice requires (or morality generally) not as a metaphysical, metaethical or logical phenomeon, but rather as a normative-semantic one. This normative-semantic difference is what interlocutors pragmatically express, namely conflicting beliefs about what the meaning of a word ought to be (in *Nationalisation* the word would be "justice"). They pragmatically express this conflict by making judgements using the word with a different meaning ot their interlocutor. This reading of certain reasonable disagreements as metalinguistic negotiations is defended, as in the case of *Spicy Soup*, on the basis of how people are disposed to carry on disagreeing despite what "justice" currently means in their community, and because they are disposed to use "justice" in systemically divergent ways.

All this means that as reasonable disagreements about metaphysical, epistemic, axiological, logical, and scientific topics increases it is likely, although not guaranteed that metalinguistic negotiations will also increase. Of course this is not guaranteed because people could, despite having wide ranging reasonable disagreements about various metaphysical, epistemic, axiological, logical, and scientific topics, simply agree that none of those topics are relevant for deliberating about justice. The dynamic between reasonable disagreements unrelated to justice (or morality generally) and metalinguistic negotiation is ultimately a contingent matter. The reasonableness of any given disagreement is separate from it being a metalinguistic negotiation, even unreasonable disagreements can also be metalinguistic negotiations. If there is a deep disagreement that hinges on a conflict about the considerations relevant for deliberating about what justice requires (or morality generally), and the people involved either do not sincerely make judgements they think others can agree to, or they fail to meet a minimum threshold of rationality (ie. they do not respond to moral reasons or make

a coherent calculation of them) their deep disagreement will be an unreasonable disagreement. Such a disagreement is still a genuine disagreement though and reading it as a metalinguistic negotiation tells us why and how it is genuine.

To move from the abstract to the concrete, reading a case like *Nationalisation* as a metalinguistic negotiation involves two suppositions. The first supposition is that their disagreement is genuine in virtue of involving a conflict in their metalinguistic beliefs about what the meaning of "justice" ought to be which they would pragmatically express, if they were to express their judgements about what justice requires. The second supposition is that Bryan and Elizabeth mean different things by "justice". Supposing those two things allows us, despite such a case not hinging on what reasonable people would literally express about what justice requires, to say that it is in fact a genuine case of disagreement.[4]

This strategy seems to hold because the reasons for those suppositions in a case of deep disagreement like *Nationalisation* are the same as it was for *Spicy Soup*. For the first supposition, it seems Bryan and Elizabeth are disposed to carry on disagreeing even when they both agree about what "justice" currently means in the community at large or according to some third-party. After all, they are not pointing to dictionaries or articles in philosophy journals to defend their moral and political judgements. At least one of them thinks those facts are irrelevant. Rather, they are likely to persist in their disagreement because they recognise that using "justice" with the meaning that their interlocutor gives it has significant normative consequences for them. The idea being that it matters what states of affairs are described as "just". This is because it will affect what normative demands are placed on them and what social arrangements will be enforced with the use of coercive political power.

For the second assumption, it seems that Bryan and Elizabeth are disposed to use "justice" in systematically divergent ways whenever they need to make judgements pertaining to matters of justice. This is because, as we have already established, they disagree about what the deliberative considerations ought to be when making moral and

[4]See Ball (2020) for someone who is sceptical of metalinguistic negotiation as a way to explain deep disagreements. This is because understanding the participants as advancing a view about what the meaning of word ought to be, either 1) entails that people are being unreasonable in trying to advance the first-order issue (eg. the economic institutions and outcomes justice requires.) by asserting a view about what the meaning of a word ought to be, or 2) fails to make sense of the reaction that a reasonable person has to have to contest the first-order claim rather than contest the view about what the meaning of word ought to be. I am not sceptical in this way. I think Ball's analysis goes wrong because he overlooks the way pragmatic expression is key to metalinguistic negotiation, and that he places an epistemic standard of reasonableness on the participants that is either irrelevant to political theory or a standard political theorist need not accept.

political judgements, and what deliberative considerations are used will determine, in some sense, how people use their moral terms across various contexts. The idea being that when people do not even agree about what deliberative considerations to use they are disposed, across various contexts, to form conflicting beliefs about what their moral terms apply to. If they then make their judgements accordingly they will *ipso facto* use their terms in systematically divergent ways. To that end, if patterns of word usage is a guide to a word's meaning, then it is plausible that in *Nationalisation* Bryan and Elizabeth mean different things by "justice". To that end, it is plausible that *Nationalisation* hinges on a pragmatically expressed conflict about what "justice" ought to mean and that it is plausible that reasonable disagreements that are deep disagreements can be read as metalinguistic negotiations.

2.2 Concept-Conception Packages

Taking stock for a moment, I have said that the idea of canonical disputes and metalinguistic negotiations gives Diverse Packages Theory the resources to read any case of reasonable disagreement as genuine. It allows the theory's explanatory model to read cases of reasonable disagreement like *Abortion* and deep disagreements like *Nationalisation* as genuine disagreements.

But of course none of this helps us explain why those reasonable disagreements occur. That is where the second moving part comes in to play. The crucial idea being that explaining why reasonable disagreements occur involves describing the differences in how people possess and use 'concept-conception packages'. The basic idea being that to explain why reasonable disagreements occur as canonical disputes or as metalinguistic negotiations we need to look at what reasonable people say (or at least what they would say) and describe the package of concepts and general beliefs about the extensions of the concept (ie. conceptions) in their minds that cause them to say those things. Doing this, I propose, begins with the thought that to explain a case of reasonable disagreement we reformulate it into a basic linguistic exchange which represents what people say or what they would say. The most general version of such an exchange for reasonable disagreements would be something like:

> Speaker 1: Society ought ϕ because ϕ-ing is what [moral-standard-x] requires.
>
> Speaker 2: Society ought to not-ϕ because not-ϕ-ing is what [moral-standard-x] requires.

In such an exchange, ϕ stands for an action or state of affairs to be realised, and

"[moral-standard-x]" being a word or words like, "justice", "promoting the good" or simply "morality". To draw out the key elements of the exchange that need to be explained, I then propose a general, and fairly uncontroversial, linguistic analysis. It first supposes that the speakers make conflicting judgements. This means that Speaker 1 and Speaker 2 say sentences that that express judgements that cannot both be true. To get a sense of what that means, we can say that each speaker makes judgements according to the types of states of affairs they believe that are the extensions of "[moral-standard-x]", namely ϕ-ing or not-ϕ-ing. They do this on the basis of the intension, or 'meaning' broadly construed, of "[moral-standard-x]" because the meaning of a term constrains the extensions of that term. We can then say that a difference between speakers' beliefs about the extensions of "[moral-standard-x]" and the meaning of "[moral-standard-x]", is what constitutes the conflicting judgements that are then expressed in their disagreement.[5]

With this basic linguistic analysis on hand, the question is how should we understand the causal process of Speaker 1 and Speaker 2 expressing conflicting judgements according to their beliefs about the extensions of "[moral-standard-x]" and meaning of "[moral-standard-x]"? I propose we understand it in terms of differences in how they possess concepts as the building blocks of thought which correspond to the meaning of terms, and how they use these concepts to form conceptions which are general beliefs about the types of states of affairs in the extensions of those concepts.[6] In short, we ought to understand it in terms of how reasonable people possess and use diverse 'concept-conception packages'.

This idea begins with the thought that for any individual speaker the meaning of "[moral-standard-x]" involves possessing the concept MORAL-STANDARD-X. This is because most if not all moral and political terms are single lexical items whose semantic content is associated with a unit of mental content which is a lexical concept. The concept MORAL-STANDARD-X is individuated by its conceptual content which is a body of information that comes in two varieties: invariable and variable.[7]

[5]Again, this is not to say anything about the *correct* or *true* meaning of "[moral-standard-x]". Rather simply to analyse what it means for Speaker 1 and Speaker 2's judgements express a conflict in mental content, with reference to their 'speaker-meaning', so to speak, of "[moral-standard-x]".

[6]The broad contours of this view, especially the distinction between concepts and conceptions is not new. See Rey (1985, 1983) for its early use in the philosophy of language and mind and Rawls (2005: 14, fn.15, 1999: 5–6, 9) for its use in political philosophy.

[7]Note, this is not the distinction Plunkett and Sundell (2015: 837–838; 2013a: 15–16) make between the "character" and "content" of a term. That distinction captures the way the meaning of words can be variant or invariant with respect to context. The distinction I make between variable and invariable content is much broader. It captures the way the content of a concept can vary between people but

The *invariable* conceptual content of MORAL-STANDARD-X does not vary between reasonable people. It remains constant and involves information about the role or function of that concept in thought. Specifically, that it ought to be used in deliberation for forming beliefs about the extensions of the concept. On the other hand, the conceptual content that does vary between reasonable people is the *variable* conceptual content of MORAL-STANDARD-X. This involves information about the morally relevant considerations to be weighed in deliberation for forming beliefs about the extensions of the concept.

When the concept MORAL-STANDARD-X is used according to its content, it provides the morally relevant considerations that reasonable people use to deliberate and form a *conception* of [moral-standard-x] which is a collection of beliefs about the types of states of affairs in their social world in the extension of MORAL-STANDARD-X. This comprises a 'concept-conception package' for "[moral-standard-x]" which is then used to make moral and political judgements. Specifically, a speaker uses "[moral-standard-x]" according to their conception of [moral-standard-x], which in turn has been formed by using their concept of MORAL-STANDARD-X.

The obvious question is why should we think of the causal process of Speaker 1 and Speaker 2 expressing conflicting judgements in this way? After all most philosophers still think of concepts as abstract entities that are analysed by coming up with definitions of terms composed of necessary and sufficient conditions, and conceptions are all the other beliefs about the objects that satisfy the conditions.[8] Such a view is at odds with what the idea of concept-conception packages supposes concepts are and how they function. To that end, my defence of the idea of concept-conceptions packages is that it is what is supported by the empirical evidence from contemporary developmental psychology on what concepts are, how they are acquired, and how they connect causally to how people use words.

In contemporary developmental psychology concepts are in the first instance taken to be bodies of information with semantic structure embedded in certain cognitive processes.[9] This body of information is taken to represent the causal and ex-

not in whether it is an actual candidate for the concept and its use in moral and political judgement making.

[8] See Margolis and Laurence (1999) for an overview of this view.

[9] Of course as with anything there is debate within the broader cognitive science whether *all* concepts are like this. For instance, Machery (2015, 2009) has argued for a broader more general idea of concepts as merely "bodies of knowledge that are used by default in the processes underlying the higher cognitive competences". See Carey (2015, 2011), Machery (2010) and Margolis and Laurence (1999) for the specifics of this debate. I do not take a stand on this broader debate, but merely pick out the view of concepts that most relates to the sort of concepts that are used in reasonable disagreements.

planatory features of states of affairs that are salient for making the inferences and predictions involved in categorising states of affairs to then form beliefs about what the concept applies to. For instance, Susan Carey (2015, 2009: Ch. 10, 13) has argued that empirical evidence on how children acquire and develop concepts shows that the type concepts involved in articulating beliefs about moral standards consist of a particular type of information structure: "intuitive theories". As Carey (2009: 361) says:

> Intuitive theories play several unique roles in mental life. These include: (1) representing causal and explanatory knowledge; (2) supporting inferences and predictions; (3) providing the current best guess concerning the essential properties of kinds, which in turn play a privileged role in categorization decisions; and (4) on some views of conceptual content, determining those aspects of conceptual role that separate meaning from belief.

The point here is that concepts have a distinct content and role. Their content involves causal, explanatory and predictive information that describes the salient properties of the states of affairs the concept applies to. Their role is to be used in categorisation decisions to form beliefs about one's social world.

These concepts are then differentiated from conceptions by the fact that empirical evidence shows that the process of conceptual change and belief revision come apart. As Carey (2009: 490, 522) says, "In some cases of knowledge acquisition we merely change our beliefs about the world; in others we change the concepts in terms of which those beliefs are composed". As Carey (2009: 522–523) argues empirical evidence shows that the former involves the typical ways that people change beliefs like discovering new evidence and "testing hypothesis that are stated in terms of already available concepts". But, the latter involves a "bootstrapping" of new placeholder information acquired by personal experiences according to the role of the concept in thought. This is how the content of a concept is determined both by facts innate to the psychology of individuals, and their social world. As Carey (2009: 522) says, "both internal conceptual role and causal connections between entities in the world and mental symbols (both social/historical causal connections and physical causal connections involving perceptual mechanisms and inferential processes) play roles in determining content."

This corresponds to the idea of concept-conceptions packages is several ways. The first way is that the way I take the content of concepts to be bodies of information that specify the morally relevant considerations for deliberating and forming beliefs about

the extensions of the concept corresponds to the content of concepts Carey describes as "intuitive" theories. The second way is that the way I take the role of concepts to involve providing the considerations to be weighed in deliberation corresponds to how Carey specifies the content of concepts are used in categorisation tasks to form beliefs (ie. conceptions). Thirdly, the way I take the content of concepts to be determined by people's innate psychological dispositions and personal experiences corresponds to how Carey specifies the bootstrapping process of concept acquisition and change. To that end, I take it that the idea of concept-conception packages as a way to understand how and why people use moral and political concepts to make judgements is supported by the best empirical evidence on what concepts are actually like and how people acquire them.

With the idea of a concept-conception package on the table we can then say that when speakers diverge in either their concepts or conceptions, they possess and use *diverse* concept-conception packages. This means that explaining the causal process of Speaker 1 and Speaker 2 expressing mutually inconsistent mental contents amounts to describing one of two sorts of facts. Either describing the facts that cause Speaker 1 and Speaker 2 to use the concept they share differently in deliberation to form divergent conceptions, or describing the facts that cause them to fix the conceptual content of some candidate concept differently, and so possess divergent concepts. All in all the basic point is that by making use of the idea of concept-conception packages we can explain *why* reasonable disagreements as a matter of speakers expressing conflict judgements occur.

2.3 The Explanatory Model

Putting the two moving parts together allows for a single explanatory model that can explain all cases of reasonable disagreement. The first moving part – reading cases as canonical disputes or metalinguistic negotiations – allows the model to correctly parse reasonable disagreements as genuine disagreements. This means it does not, like theories that focus purely on how concepts are *used*, declare deep disagreements as verbal disagreements. It can correctly parse such disagreements as genuine by reading them as metalinguistic negotiations. Which means that they hinge on pragmatically expressed conflicting beliefs about what the meaning of moral or political term ought to be. In addition it can also correctly parse ordinary reasonable disagreements as genuine by reading them as canonical disputes. Which means that they hinge on a literally expressed conflict judgements about the extensions of moral or political term.

The second moving part – the idea of Concept-Conception packages – allows the

model to then *causally* explain why reasonable disagreements, once classified as canonical disputes or metalinguistic negotiations, occur. When a given case of reasonable disagreement is read as a canonical dispute, the model cites the facts that cause reasonable people to diverge in their conceptions. These will be facts about the way reasonable people use their concept to make their categorisation decisions in slightly different ways to then form conflicting beliefs about the extensions of the concept. This is because a canonical dispute is a disagreement that hinges on a conflict about what is literally expressed. What is literally expressed in those disagreements are people's conceptions which are their general beliefs about the extensions of the concept they are using. In such a case the model assumes that disagreeing parties share a concept, but end up possessing and using diverse 'concept-conception packages' by diverging in their conceptions.

On the other hand, when a given case of reasonable disagreement is read as a metalinguistic negotiation, the model first cites the facts that cause reasonable people to diverge in their concepts and because of this divergence how they diverge in their conceptions as well. These will be facts about the conflicting information reasonable people use to make their categorisation decisions (because they possess conflicting concepts) and then in turn form conflicting beliefs about the extensions of the concept. This is because a metalinguistic negotiation is a disagreement that hinges on a conflict about what the meaning of a word ought to be which is pragmatically expressed by their judgements. People's possession and use of divergent concepts pragmatically expresses their tacit beliefs about what the meaning of a word ought to be. In such a case, the model supposes that disagreeing parties share neither concepts nor conceptions, and so end up possessing and using diverse 'concept-conception packages' when making moral and political judgements.

3 Applying the Model

As I have said, the point of this chapter is to argue for and defend Diverse Packages Theory as the *best* explanation of reasonable disagreement about justice. This will involve two lines of argument. The first is that we have good reason to think it can explain reasonable disagreements in the way I have claimed. This will involve applying the model to the cases that motivate it and describing in more detail how the model exactly works in the way I have summarised so far. This will explain precisely what it means for reasonable people to diverge in either their concepts or conceptions.

The second line of argument is that Diverse Packages Theory has comparative ad-

vantages over extant explanations of reasonable disagreement. This will involve showing how it avoids the problems that blight the extant theories and still offers a more explanatorily powerful and more parsimonious explanation. This section takes up the first argument whilst the next section takes up the second.

3.1 Canonical Disputes and Divergent Conceptions

As I have said, one way Diverse Packages Theory's explanatory model explains why cases of reasonable disagreement occur is by reading them as canonical disputes, and describing how people possess and use diverse concept-conception packages by diverging in the conception part of that package. To understand precisely what this means consider again the sort of reasonable disagreement that motivates it:

> *Abortion:* Barry and Nora are discussing the laws concerning abortion that their society ought to enact. Barry judges their society ought to enact laws that permit abortion in the first trimester because he believes respecting women's bodily autonomy is what justice requires. This is because respecting women's bodily autonomy outweighs the value of human life in a fetus. Nora, on the other hand, judges their society ought to outlaw abortion in the first trimester barring exceptional circumstances because she believes preserving the value of human life is what justice requires. This is because the instrumental and non-instrumental value of human life outweighs women's bodily autonomy.

In keeping with the explanatory model we reformulate the disagreement into a basic linguistic exchange:

> Barry: We ought to enact laws that permit abortion in the first trimester because respecting women's bodily autonomy is what justice requires.
>
> Nora: No, we ought to outlaw abortion in the first trimester barring exceptional circumstances because preserving the value of human life is what justice requires.
>
> Barry: No it doesn't, respecting women's bodily autonomy outweighs the value of human life in a fetus.
>
> Nora: Yes it does, the instrumental and non-instrumental value of human life outweighs women's bodily autonomy.

Given what Barry and Nora say to each other, namely sentences that express conflicting beliefs about the abortion laws justice requires, the model begins by reading *Abortion* as a canonical dispute. This means first supposing that that their disagreement involves conflicting judgements about the abortion laws justice requires that are literally expressed. This is justified by the evidence of what Barry and Nora say to each other. Second, it means assuming that Barry and Nora mean the same things by their words. This is because otherwise they could not literally express conflicting judgements. They would be literally expressing judgements with different truth-conditions, and therefore not having a disagreement at all.

The model then proposes that since the meaning of "justice" involves possessing the concept JUSTICE, Barry and Nora must also share the concept JUSTICE. It then proposes that given their disagreement is genuine in virtue of a conflict in their literally expressed judgements about what justice requires, they must have conflicting general beliefs about the extensions of JUSTICE. This is because people make the literally expressed judgements according to those general beliefs. This amounts to saying that Barry and Nora diverge in their *conceptions* of justice because conceptions of justice are merely beliefs about 'the right distribution of rights, opportunities and resources amongst people, institutions and social systems'.[10] In simpler terms, how the speakers conflict about what they each take to be the *right* generalisation of what justice requires. Importantly, these conflicting beliefs can be either implicit or explicit beliefs. If the beliefs are implicit they will likely simply be an array of case-specific judgements. If the beliefs are explicit the representations will likely be specific normative principles.

The model then explains why Barry and Nora have divergent conceptions of justice despite sharing the concept JUSTICE by citing differences in their dispositions for weighing deliberative considerations. It proposes first that the deliberative considerations for forming conceptions of justice are provided by the conceptual content that individuates the concept JUSTICE and that the considerations are shared because the parties share the concept. This, as we have seen, is in keeping with the empirical evidence from developmental psychology on what concepts are and their role in thought. The evidence shows that concepts are individuated by their content

[10] This explication of a *conception* of justice summarises what I take Rawls (1999: 5–6, 9, 54) to be talking about in various places as the "proper distribution of the benefits and burdens of social cooperation", "the appropriate distributive shares" and the point of his own conception of justice being to describe the right distribution of the chief primary goods – "rights, liberties, and opportunities, and income and wealth" – that the basic structure of society – "the political constitution and the principal economic and social arrangements" – *can* dispense.

which is a body of information that people use to make categorisation decisions and then on the basis of those decisions form beliefs about what the concept applies to. This corresponds to the casual process of reasonable people's shared concept of JUSTICE having content which reasonable people use to categorise their social world into morally relevant considerations. When these considerations are weighed in deliberation they allow reasonable people to form beliefs about what justice requires. When speakers share a concept, the information that determines or is inputted into their deliberations for forming conceptions is the same.

The model proposes that Barry and Nora form divergent conceptions because their deliberative process is affected by psychological dispositions that are either innate or acquired by personal experiences. Differences in reasonable people's psychological makeup or personal experiences mean they have dispositions to assign different weights to their various deliberative considerations.

In actual cases of course it would be an empirical matter what the specific deliberative considerations in play are, and what the dispositions and personal experiences that affect the assignment of weights are. But with *Abortion*, the case itself provides some evidence of what they could be. Barry and Nora clearly see that people's autonomy, and the value of human life are the relevant deliberative considerations to be weighing up when forming conceptions of justice. But something in Barry's personal experiences, or innate psychological make up means he has a disposition to assign greater weight to the value of autonomy achieved by permitting abortion over the value of human life. Perhaps he has been exposed through childhood to cases where adults have not been able to exercise their bodily autonomy.

Nora, in contrast to Barry, is disposed to assign the considerations in exactly the opposite way. Perhaps she has been exposed to cases where mothers and newborns live happy lives that benefit themselves and their wider community. Whatever the precise source of their dispositions the point is that their deliberative processes for forming conceptions of justice are affected by their psychological dispositions to assign different weights to their shared deliberative considerations. As a result, they possess diverse 'concept-conception packages' by diverging in their conceptions.

When Barry and Nora use their diverse concept-conception packages they end up making intractably conflicting political judgements which comprise *Abortion*. The idea being that when they use the term "justice" according to their conception of justice, they make conflicting judgements about the institutions and outcomes justice requires. This is because their conceptions of justice are their general beliefs about the extensions of JUSTICE. As such, when they use "justice" according to these con-

flicting beliefs, they will end up making judgements that conflict which they then literally express.

Their judgements will conflict *intractably* because whether their dispositions are innate or acquired by personal experiences, they cannot be altered or erased by simply rehearsing their deliberation with others. Innate dispositions are after all innate, and personal experiences are events in the past that cannot be changed. However, people end up deliberating it will be caused by their sum of innate dispositions and personal experiences interacting in particular ways. This does not of course mean that deliberating with others can have no effect. Barry and Nora's discussion might itself become a new shared personal experience that compensates for their innate dispositions and pushes them to agree. But given particular differences in dispositions and personal experiences, *those* differences are what make the difference between Barry and Nora having a reasonable disagreement and coming to a reasonable agreement.

In sum, one argument for Diverse Packages Theory is that its explanatory model can explain why cases of reasonable disagreement like *Abortion* occur by reading them as canonical disputes. By treating them as canonical disputes it correctly interprets them as genuine disagreements in virtue of hinging on conflicting judgements that are literally expressed by the disagreeing parties. This allows the model to explain them by showing how people possess and use diverse concept-conception packages when making moral and political judgements. It describes the dispositions and personal experiences that affect how reasonable people weigh their shared moral considerations, which causes them to form divergent conceptions of justice. When reasonable people then go on to use their divergent concept-conception packages they end up making conflicting moral and political judgements and so end up in reasonable disagreements rather than reasonable agreements.

3.2 Metalinguistic Negotiations and Divergent Concepts

As I have said already, Diverse Packages Theory's explanatory model can also explain why cases of reasonable disagreement occur by reading them as metalinguistic negotiations and describing how people possess and use diverse concept-conception packages, by diverging in their *concepts*. To understand better what that means consider the type of case that motivated it, namely a reasonable disagreement that is a deep disagreement:

> *Nationalisation:* Bryan and Elizabeth are discussing the economic structure their society ought to have. Elizabeth judges their society ought to nationalise, at the very least, some key industries because a society's

> productive capacity being for the mutual benefit of all is what justice requires. Bryan judges their society ought not to nationalise any industries because protecting people's natural moral right to their body and private property is what justice requires. This is because he believes natural moral rights are what matter, benefiting everyone with society's productive capacity is irrelevant and has nothing to do with what justice requires. On the contrary, Elizabeth believes materially benefiting people is what matters, natural rights to private property are irrelevant and have nothing to do with what justice requires.

In keeping with the explanatory model we reformulate the disagreement into a basic linguistic exchange to get the following:

> Elizabeth: We ought to nationalise, at the very least, some key industries because a society's productive capacity being beneficial to all is what justice requires.
>
> Bryan: No, we ought not to nationalise any industries because protecting people's natural moral right to their body and private property is what justice requires.
>
> Elizabeth: No we shouldn't, benefiting everyone is what matters, natural rights to private property is irrelevant for justice.
>
> Bryan: Yes, we should, natural rights are what matter, benefiting everyone with society's productive capacity is irrelevant for justice.

As we saw from §2.1 the best way to understand how and why a case like *Nationalisation* is a genuine disagreement is to read it as a metalinguistic negotiation. This means the model supposes that Bryan and Elizabeth's disagreement involves a conflict over what they believe the meaning of "justice" ought to be which is pragmatically expressed by what they say to each other, and that as result they mean different things by "justice". What justifies these suppositions is that, again as we saw in §2.1, Bryan and Elizabeth are disposed to use "justice" in systematically divergent ways, and they are disposed to carry on disagreeing even when they both agree about what "justice" currently means in the community at large or according to some third-party. These two bits of evidence from Bryan and Elizabeth's exchange justifies reading the case as a metalinguistic negotiation.

The model then supposes that given the meaning of "justice" involves possessing the concept JUSTICE, this means Bryan and Elizabeth diverge in their concept of JUSTICE. As a result, it supposes that their disagreement is genuine in virtue of their

what they say to each other pragmatically expressing a conflict over what they believe the content of JUSTICE ought to be. As such, the model then explains why Bryan and Elizabeth hold such conflicting metalinguistic beliefs by explaining how Bryan and Elizabeth possess and use diverse concept-conception packages by diverging in the concept of JUSTICE they possess and use. After all possessing and using divergent concepts is what pragmatically expresses one's beliefs about what the content of the concepts ought to be.

Since concepts are individuated by their conceptual content, explaining why Bryan and Elizabeth possess and use divergent concepts of JUSTICE amounts to showing how they can possess different conceptual content for the concept JUSTICE. To avoid making *Nationalisation* a verbal disagreement this will be a matter of showing how Bryan and Elizabeth only possess different variable content for JUSTICE, but share the invariable content.

We can say that for the concept JUSTICE the invariable content that Bryan and Elizabeth share amounts to a description like, 'to form beliefs about the right or morally correct distribution of rights, opportunities and resources amongst people, institutions and social systems'.[11] In short, it ammounts to something like 'to form the sorts of beliefs that are part of conceptions of justice'. This is because, as I have said, the invariable content of a concept contains information that specifies the role of the concept in thought, ie. its use in forming general beliefs about the extensions of the concept. Given 'distributions of rights, opportunities and resources amongst people, institutions and social systems' are the types of states affairs that are in the extension of JUSTICE, the role of the concept is to form beliefs about the distributions that are actually in its extension, namely the right or *morally correct* distributions.

The variable content on the other hand will be a description of the considerations morally salient for forming those beliefs (ie. conception of justice). As the linguistic exchange makes plain, for Bryan the variable content will be something like 'natural moral rights to their body and private property are the morally relevant considerations for justice'. For Elizabeth it will be something like, 'productive capacity being for the mutual benefit of all are the morally relevant considerations for justice'. In real world cases, what these contents actually are will largely be an empirical matter. But, for now it suffices that what it means for Bryan and Elizabeth to possess diver-

[11] This explication of the invariable content of JUSTICE follows Rawls's (1999: 5–6, 9) distinction between the concept of justice and conceptions of justice. The former describes the common function of the latter. The explication of the conceptual content of justice here merely follows this idea with respect to how I have so far described conceptions of justice as describing 'the *correct* distribution of rights, opportunities and resources amongst people, institutions and social systems'.

gent concepts of JUSTICE is that they possess conflicting variable conceptual content for JUSTICE. Therefore, what it means to explain why Bryan and Elizabeth diverge in their concepts amounts to explaining why they possess this conflicting conceptual content.

The model explains why Bryan and Elizabeth possess conflicting conceptual content by citing differences in the facts that cause them to fix the content of JUSTICE in conflicting ways. The question then is, differences in what facts cause reasonable people to fix the content of JUSTICE in conflicting ways? The explanatory model cites differences in psychological dispositions that are innate or acquired by personal experience. These psychological dispositions affect the cognitive process by which Bryan and Elizabeth acquire concepts and slowly fix their content over time. Different dispositions will result in reasonable people either acquiring slightly different concepts of JUSTICE or fixing the content of their existing concept of JUSTICE in slightly different ways over time.

Note, the use of "psychological dispositions" here is deliberately broad. This is because Diverse Packages Theory is neutral with respect to on-going debates in cognitive science and developmental psychology about how concepts are precisely acquired and how their content is fixed. Diverse Packages theory is neutral on whether distinct concepts are acquired, by distinct personal experiences interacting with general learning mechanisms in one's cognitive architecture, by distinct innate learning mechanisms that affect how experiences are taken as inputs for forming concepts, or by innate core cognitive mechanisms interacting with innate conceptual primitives to construct new concepts as a response to one's personal experiences.[12] These are all live options for Diverse Packages Theory since the notion of a psychological dispositions can capture the general causal capacities for individuals to acquire concepts in all those options. A philosophical explanation need not take sides in this empirical debate.

With all that in mind, the model proposes that differences in Bryan and Elizabeth's innate and acquired psychological dispositions cause them to possess and use divergent concepts of JUSTICE. Although citing any particular dispositions is a matter of empirical investigation, we can make certain guesses in *Nationalisation* to demonstrate how the model works. For instance, Bryan is perhaps disposed to fixing the content of his concept of JUSTICE according to his early experiences reading books on political economy that emphasise people's natural moral rights. As such he fixes the content of JUSTICE in a way that ensures "the protection of people's natural moral

[12] See Carey (2015, 2011, 2009: Ch. 11) Hamlin (2015), Kalish (2015) and Hampton (2015) for a good overview of these various positions.

rights" as the only relevant consideration for justice. Elizabeth on the other hand fixes the content of her concept of JUSTICE, perhaps according to her innate disposition towards benefiting everyone equally, and perhaps her early experiences living in an unusually egalitarian household. As such, she fixes the content of JUSTICE in a way that ensures "equal beneficence to individuals" as the only relevant consideration for justice. As I have said the notion of innate and acquired psychological dispositions here is deliberately broad to include a wide range of concept acquisition and concept change mechanisms. The important point is that Bryan and Elizabeth fix the content of their concepts in different ways and therefore possess divergent concepts of JUSTICE.

The model then proposes that given Bryan and Elizabeth possess divergent concepts of JUSTICE, they will form divergent conceptions of justice. This is because when they use these divergent concepts they will provide conflicting sets of deliberative considerations for forming conceptions of justice. As a result, let alone how they weigh such considerations, the very fact they will be weighing different deliberative considerations will result in Bryan and Elizabeth forming divergent conceptions of justice. At this point, the model has once again explained how reasonable people, Bryan and Elizabeth in this case, possess diverse concept-conception packages.

When Bryan and Elizabeth use their diverse concept-conception packages to make their political judgements, they will make intractably conflicting political judgements about what justice requires. This is because they make their political judgements according to their conceptions of justice. But, importantly these judgements do not conflict in virtue of being contents that cannot both be true. After all, the truth-conditions for Bryan and Elizabeth's judgements are different. Rather the judgements Bryan and Elizabeth make conflict in virtue of what they would *pragmatically* express. They would pragmatically express what they each believe the concept of JUSTICE ought to be for making judgements about what justice requires. Given, the concept JUSTICE constitutes the meaning of "justice", this then explains how reasonable people have a metalinguistic negotiation about what the meaning of "justice" ought to be.

This conflict in political judgements is intractable because what causes it are differences in dispositions that cannot be altered or erased by Bryan and Elizabeth rehearsing their deliberative process with each other. They cannot alter their innate dispositions or the personal experiences that make them acquire the dispositions to fix the content of their concepts in conflicting ways. Deliberation with others, as it usually occurs, involves exchanging reasons for endorsing a particular conception of

justice. Rehearsing one's deliberative process for forming conceptions of justice with those that disagree is not going to affect the conflict over what the concept of JUSTICE ought to be.

But of course none of this is to disregard the attempt to change people's concepts. After all, one prominent research project in contemporary philosophy attempts to do just this. For example, Conceptual Engineering and Conceptual Ethics are research projects that have recently emerged with an aim to revise or replace the concepts we use.[13] But, to say that Bryan and Elizabeth make intractably conflicting political judgements because they use divergent concepts is merely to recognise that the only way Bryan and Elizabeth's disagreement can be resolved is if at least one of them changes their concept. The only way to effectively alter the content of one's concepts is for the deliberation with others about the content of JUSTICE to itself become an experience that disposes them to change the content of their concept. But, this is a far cry from deliberating with one's interlocutor about much weight one is giving to a certain piece of evidence or rehearsing whether one is applying a rule of inference correctly. Insofar as the deliberation with others involves discussing the weights of various consideration and trying to explain why some consideration is decisive for forming a particular conception of justice it will not make them agree. Rather, they must deliberate about how to correctly fix their concepts.

With all that, the model has explained how Bryan and Elizabeth possessing and using diverse concept-conception packages causes them to have a reasonable disagreement rather than reasonable agreement. Bryan and Elizabeth's dispositions cause then to possess divergent concepts, which cause them to form divergent conceptions. When they use their respective concept-conception packages to make their judgements they lead them to make judgements that conflict in virtue of pragmatically expressing a conflict about what the meaning of "justice" ought to be for making political judgements about what justice requires.

But of course *Nationalisation* is not representative of all forms of deep disagreement. Deep disagreements can come in roughly two varieties: cases of *Direct Deep Disagreements* and cases of *Indirect Deep Disagreements*. *Direct Deep Disagreements* are cases where reasonable people disagree directly about what the deliberative considerations ought to be for deliberating about some moral or political topic. In *Nationalisation* this topic was justice. But, *Indirect Deep Disagreements* are not like this. They are cases where reasonable people disagree *indirectly* about what the deliberative

[13]See Burgess and Plunkett (2013a,b) on Conceptual Ethics and Cappelen and Plunkett (2020) on Conceptual Engineering for an overview and examples of philosophical projects that involve explicitly changing people's concepts in society by trying to persuade people.

considerations for deliberating about a particular moral or political topic ought to be. Reasonable people in these cases disagree about the deliberative considerations by way of disagreeing about what the content of a particular deliberative consideration ought to be.

So far I have only explicitly argued for how Diverse Packages Theory's model can explain reasonable disagreements that are cases of *Direct Deep Disagreement*. But, I submit, the model can also explain cases of *Indirect Deep Disagreement*. To see how consider the following case that is like *Abortion* but slightly different:

> *Indirect Abortion:* Barry and Nora are discussing the abortion laws their society ought to enact. Nora judges their society ought to outlaw abortion because she believes, although it does restrict women's autonomy, fetuses are innocent persons which means killing them is murder and laws against murder are what justice requires. Barry, on the other hand, judges their society ought to enact laws that permit abortion in the first trimester because, whilst he agrees justice requires laws against murder, he believes fetuses in the first trimester are not people, so killing them is not murder and so permitting abortion to protect women's autonomy is what justice requires.

Indirect Abortion is an *Indirect Deep Disagreement* because Barry and Nora disagree *indirectly* over what the deliberative considerations ought to be for deliberating about what justice requires. They disagree over what the content of a particular deliberative consideration – moral personhood – ought to be for determining what justice requires. This is despite them agreeing that moral personhood is a relevant consideration for deliberating about what justice requires.

To explain *Indirect Abortion*, we begin by reformulating it into a basic linguistic exchange:

> Nora: We ought to outlaw abortion in the first trimester because, although it does restrict women's autonomy, fetuses are moral persons which means killing them is murder and laws against murder are what justice requires.
>
> Barry: We ought to protect the right to an abortion in the first trimester because, whilst he agrees justice requires laws against murder, a fetus is not a moral person, so respecting women's bodily autonomy is what justice requires.

As with *Nationalisation* the model would read *Indirect Abortion* as a metalinguistic negotiation. But, this time it does not suppose that the disagreeing parties mean different things by "justice" or that their disagreement involves a conflict over that the meaning of "justice" ought to be. Rather it assumes that Barry and Nora mean different things by "moral person", and that their disagreement hinges on a conflict over what the meaning of "moral person" ought to be. This is because in *Indirect Abortion*, Barry and Nora disagree about what the deliberative considerations ought to be for deliberating about what justice requires, by way of disagreeing about the content of one of the considerations. That consideration is "moral person".

The model would then explain *Indirect Abortion*, in much the same way it explained *Nationalisation* except instead of focusing on the concept JUSTICE, it would focus on the concept MORAL PERSON. It would explain how Barry and Nora diverge in their concept-conception package by explaining how they diverge in their concept of MORAL PERSON and then how this divergence causes them to forming conflicting conceptions of justice. For the sake of repetition I will not retrace the steps the model takes for explaining all that. Rather it will suffice to say that it would explain how Barry and Nora diverge in their concepts of MORAL PERSON by citing the innate and acquired dispositions that affect how they fix the content of the concept. To that end, the model would also be able to explain reasonable disagreements that are *Indirect Deep Disagreements*.

This concludes the argument for Diverse Packages Theory on independent grounds. I have shown how its explanatory model can explain the cases of deep disagreement that motivated it. It reads them as metalinguistic negotiations. It supposes the reasonable people in such disagreements mean something different by at least one of their words, and that their disagreement hinges on a pragmatically expressed conflict about what the meaning of that word ought to be. The model then explains why such reasonable disagreements occur by citing the facts that cause people to possess and use diverse concept-conception packages that diverge in the concept. The model shows that when reasonable people use concept-conception packages that diverge in this way to make moral and political judgements, they pragmatically express a conflict about what the meaning of a word ought to be. As such, they end up in reasonable disagreements rather than reasonable agreements.

4 Comparative Advantages

In addition to the independent grounds for endorsing Diverse Packages Theory there are two comparative grounds. The first ground is that it is more explanatorily powerful than many extant theories because of how it can explain cases of deep disagreement. The second ground is that it relies on a more parsimonious view of concepts when compared to the only other theory that matches it for explanatory power, namely Dworkin's theory of interpretive concepts. In what follows I detail both advantages.

4.1 Explanatory Power

As we saw in Chapter 2 many theorists, seeking to develop Rawls's idea of the "burdens of judgement" further, propose theories that explain reasonable disagreement by citing facts purely about how people use concepts. These theories cite differences in reasonable people's dispositions and personal experiences as the facts that cause reasonable people to use their concepts to form conflicting conceptions. Although the details of each particular theory is slightly different, the underlying strategy is the same. They explain how reasonable people systematically diverge in using their moral terms to make conflict political judgements by describing how they diverge in using their shared concepts to form conflicting conceptions.

The problem with these theories, as I argued in Chapter 2, is that they face a dilemma. On one horn they can accept that deep disagreements are genuine disagreements, in which case they do not have the resources to explain deep disagreements as genuine disagreements and therefore have a serious explanatory weakness. On the other horn, they can deny that deep disagreements are genuine and that therefore they need not be explained. But, this is deeply counterintuitive. Reasonable disagreements which are deep disagreements seem to involve a substantive normative disagreement, and the participants seem to think they are worth having because how the disagreement is resolved matters for how coercive political power would be used.

But Diverse Packages Theory can explain deep disagreements as genuine disagreements. This is because it allows for the possibility that reasonable disagreements may be metalinguistic negotiations. This means they can sometimes hinge on pragmatically expressed conflicting beliefs about what a moral term ought to mean. To explain such disagreements, Diverse Packages Theory cites facts about both concept use *and* possession. This is what the idea of "concept-conception packages" allows for. It allows for the explanation of reasonable disagreements as both canonical disputes and

as metalinguistic negotiations in a single explanatory model.

4.2 Parsimony

The most significant advantage of Diverse Packages Theory is that its explanatory model relies on a more parsimonious view of concepts than Dworkin's theory of interpretive concepts. This is significant because Dworkin's theory is the only other extant theory that can explain reasonable disagreements that are deep disagreements. By relying on a more parsimonious theory of concepts it avoids the problems Dworkin's theory faces.

Dworkin's theory, as I argued in Chapter 2, proposes an entirely new type of concept – an interpretive concept – to explain reasonable disagreements.[14] I argued that although Dworkin's theory can explain both ordinary reasonable disagreements and deep disagreements, it faces two problems: the Virtue of Reasonableness Problem, and Regression of Interpretation Problem. The first problem is that in declaring all moral and political concepts as interpretative, Dworkin's theory makes reasonableness itself a matter of reasonable disagreement. The second problem is that when reasonable people in deep disagreements try to communicate why they make their judgements and the content of the concept they think ought to be used, Dworkin's theory entails that they will descend into a regression of interpretation.

Both issues stem from the way Dworkin thinks interpretive concepts work and his global distinction on all moral and political concepts being interpretive. Diverse Packages Theory avoids both problems because it offers a far more parsimonious view of concepts and how they work. It neither posits an entirely new type of concept that is unsupported by empirical data, or supposes that all moral and political concepts are inextricably interconnected. Rather, it merely posits a new type of genuine disagreement.[15] This leaves concepts as they are without the need to declare all moral and political concepts as interpretive, or that the content of any single one of them depends on a host of other interpretive concepts.

5 Objections

Taking stock for a moment, so far I have argued for Diverse Packages Theory as the *best* explanation of reasonable disagreement on the basis of two lines of argument. The

[14] See also Plunkett and Sundell (2013b: 251–252) for this broad understanding of Dworkin's theory.

[15] See Plunkett and Sundell (2013b) for this specific advantage of metalinguistic negotiation over Dworkin's theory when it comes to legal disagreement.

first argument is that it can make good on what motivates it and so actually explain cases of reasonable disagreement that are both ordinary disagreements and deep disagreements. The second argument is that it is better than other extant explanations of reasonable disagreement on the metrics of explanatory power and parsimony. It can explain cases of deep disagreement when other theories cannot, and it does so without committing to an entirely new theory of concepts. The underlying strategy has been that by reading cases of deep disagreement as metalinguistic negotiations, we can vindicate the intuition that they are genuine disagreements and that they can be explained alongside ordinary cases of reasonable disagreement in a single explanatory model.

I want to now consider three objections that might be raised against this strategy, namely that in employing the idea of metalinguistic negotiations, Diverse Packages Theory has inadvertent consequences for reasonable disagreement. The first concerns externalism about meaning. The second concerns the pointlessness of verbal disputes. The third concerns topic discontinuity.

5.1 Semantic Externalism

In arguing for Diverse Packages Theory I presented *Nationalisation* and *Indirect Abortion* as cases of disagreement where I claimed that reasonable people's words can express different contents. In simpler terms, that sometimes when reasonable people use words I have claimed they mean different things by them. But, one might think this relies on a controversial internalist metasemantic theory. This means the model assumes that what disagreeing parties mean by their words and therefore the conceptual content of their concepts are determined by facts internal to the disagreeing parties, namely their psychological dispositions. But, so the objection goes, this is not how the meaning of words and the content of concepts is determined. The meaning of our words is determined by facts external to the speaker.[16] It does not matter what our dispositions are like. That is irrelevant to the content our words express when we use them. What is relevant are facts external to us about how the community at large uses the word or how a word was first used to refer to some part of the world. The upshot of this is that one of the key moving parts of Diverse Packages theory is at best controversial, and at worst false.

I have two responses to this sort of objection. One clarificatory, the other methodological. The clarificatory response is that in everything I have said so far, I have left it open whether the psychological dispositions that affect how reasonable people fix

[16] See Sawyer (2020) and Cappelen (2018: Ch. 6) for this sort of view.

the variable content of their concepts are innate or acquired by personal experience. It could well be that all the relevant psychological dispositions are acquired by personal experiences. This would mean that ultimately facts external to speakers are what cause speakers to fix the conceptual content of their concepts differently. Moreover, this is not by accident. It is a consequence of taking seriously the empirical evidence on concepts, which shows that both facts internal and external to an individual determine the meaning of the words they use. As such, Diverse Packages Theory takes no stand on how many of our psychological dispositions relating to content fixing are innate or themselves created by facts external to the speaker. Therefore, the externalist could well be correct on this front and still accept Diverse Packages Theory.

The methodological response is that ultimately Diverse Packages Theory is for political theorists to solve a problem in political theory, namely to explain reasonable disagreement. As such, the efficacy of Diverse Packages Theory is independent of debates in the philosophy of language about which psychological dispositions track the *correct* way to fix the content of concepts. Diverse Packages Theory is entirely consistent with an externalist metasemantic theory that says the conceptual content of our terms is determined by specific facts external to the speaker irrespective of their dispositions or metalinguistic beliefs. In such cases, the externalist has merely become a participant in a metalinguistic negotiation. As such, Diverse Packages Theory leaves it open for an externalist to still say that in a metalinguistic negotiation, at least one of the parties is mistaken about what their words mean because they have fixed the content of their concepts incorrectly. Nothing in Diverse Packages Theory precludes this, but it is irrelevant to evaluating whether Diverse Packages Theory does the work it is supposed to do for the political theorist's problem of explaining reasonable disagreement. Of course whether the externalist's metasemantic theory is relevant for the political theorist's problem is an open question. But, *that* question is part of an orthogonal methodological debate about how different parts of philosophy are related, and so is unrelated to the question here and now of whether Diverse Packages Theory is a good theory to solve the political theorist's problem.

5.2 Pointless Verbal Disputes

In arguing for Diverse Packages Theory's model I said that one motivation for it was that in being able to read cases of reasonable disagreement as metalinguistic negotiations it can vindicate the intuition that deep disagreements are genuine. By assuming that they hinge on a pragmatically expressed conflict about what the meaning of a moral term ought to be we can understand how they are genuine despite disagreeing

parties not literally expressing conflicting beliefs by their judgements.

An objection one might have, following Chalmers (2011: 522–525), is that Diverse Packages Theory makes deep disagreements into pointless verbal disputes. The core idea being that even if there is a sense in which deep disagreements, when read as metalinguistic negotiations, are genuine in virtue of hinging on a conflict in what concept ought to be used, they are pointless to the first-order practical matter that is expressed by reasonable people's political judgements. And, as Chalmers (2011: 525) says, this sort of pointlessness is a heuristic guide to the presence of verbal disputes. As such, using metalinguistic negotiation to vindicate deep disagreements as genuine disagreements was in vain. They are pointless verbal disputes anyway and pointless disputes are no help for political theorists who hope to show reasonable disagreements were worth explaining.

To see how this objection works consider *Nationalisation* again. In that case, the thought would be that when it is read as a metalinguistic negotiation it turns out to hinge not on the first-order disagreement we thought it did. It does not hinge on the economic system Bryan and Elizabeth believe justice requires. Rather, it hinges on a conflict about what the concept of JUSTICE ought to be. But, insofar as this generates the first-order disagreement, it is pointless to it because all it involves is a conflict about what considerations are relevant for deliberating about justice. And, resolving that dispute will resolve the first-order disagreement. But, this resolution doesn't come about because Bryan and Elizabeth settled something about justice. Rather it comes about because they settle on which considerations they ought to be weighing up to form beliefs about what justice requires.

But, I submit, this objection goes wrong for two reasons. The first is that it is plainly wrong to conclude that metalinguistic negotiations are pointless with respect to the first-order moral and political issues that reasonable people are disagree about. The resolution of a metalinguistic negotiation about what concept ought to be used has significant normative consequences when they involve the kinds of concepts that usually feature in reasonable disagreements. Reasonable disagreements as should be plain from examples like *Abortion*, *Nationalisation*, and *Indirect Abortion*, can potentially involve concepts like JUSTICE, MORAL GOODNESS, MORALLY RIGHT, MORAL PERSON and many other moral concepts. These are by their very nature normative concepts and therefore if people diverge over them and decide to resolve the divergence, they will have serious and pervasive normative consequences. For instance, that some states of affairs is picked out as required by JUSTICE is typically taken to justify realising it with the use of coercive political power. As such people will be forced

to comply with laws that enforce a state of affairs that counts as just according to their interlocutor's concept. One can imagine how this will play out in a similar way for a whole host of reasonable disagreements that feature moral and political concepts. To that end, deep disagreements read as metalinguistic negotiations are not pointless, and as such not necessarily or heuristically verbal disputes.

For a non-political example, consider a reasonable disagreement that is centred on a topic in interpersonal morality, "the good life". Let us suppose that it is a metalinguistic negotiation and as Chalmers supposes parties resolve to use the term "good life" to mean 'lives spent satisfying selfish desires' rather than to use "good life" to mean 'lives spent helping others'. Clearly this will have profound normative differences for these individuals. It will affect what they see as demanded of them by morality and therefore how they should lead their lives.

The second reason the objection goes wrong is because Diverse Packages Theory already has a definition of what verbal disputes are and so Chalmers's way of identifying them is unnecessary. As I argued, people will have a verbal disagreement when they do not share the invariable content of a concept. This is because it would show they do not possess diverging concepts that would count as competing candidates. Rather, they would be distinct concepts altogether who play completely different roles and having completely different functions for those who wish to use them. A disagreement using those concepts would be verbal unless we decomposed it further into disagreement about some more fundamental concepts that people shared, or some disagreement about entire conceptual schemes.[17] Barring anything like that, Diverse Packages Theory has already provided an analysis of when reasonable disagreements are verbal disputes.

5.3 Topic Discontinuity

Another objection against Diverse Packages Theory might be that, even if we accept that understanding deep disagreements as metalinguistic negotiations does not entail they are pointless verbal disputes, it might still entail they involve a form of topic discontinuity. That is to say that the model shows that deep disagreements are cases where reasonable people are trying to change the topic of their discussion, rather than trying to engage in sustained disagreement about the same topic.

The objection begins with the thought that concepts play a central role in picking

[17] See Chalmers (2011: 548–563) on bedrock concepts and bedrock disputes, and Midgley (1992) on conceptual schemes.

out the "topic" or general subject matter of thought.[18] For instance, when someone is deciding to buy a house and reflecting about what they ought to do, a number of thoughts will typically run through their head. They will think about various things relating to their finances, various things relating to the condition of the house, perhaps what others think about the house or their finances as well. But all through this, their thoughts centre on a particular topic picked out by a concept, or more likely a cluster of concepts, relating to 'purchasing a house'. This will be the same for people engaged in a discussion or disagreement. A sameness in concepts will guarantee a sameness in topic.

But, so the objection goes, if Diverse Packages Theory is right that reasonable disagreements like *Nationalisation* and *Indirect Abortion* are metalinguistic negotiations, then it shows that the people in those disagreements are not actually talking about the same topic. This is because as Diverse Packages Theory explains those cases, the disagreeing parties are each making a judgement according to their divergent concepts. This is not a pointless verbal dispute, but it does indicate that the disagreeing parties are trying to have a discussion centred on different topics, namely topics individuated by the concepts they are each individually using. This would be to say that in *Nationalisation*, Bryan is trying to have a discussion about what justice requires as it pertains to his concept of JUSTICE, but Elizabeth is trying to have discussion about what justice requires as it pertains to her concept. But, neither of them realise that they are each trying to change the topic and so there is no sustained disagreement on a single topic of discussion.

But this objection is unwarranted because Diverse Packages Theory is fundamentally neutral with respect to a wide variety of views on topic continuity. For instance, it is compatible with the sort of view argued for by Schroeter and Schroeter (2014: 12–16), and Sawyer (2020: 385–390, 2018b: 10–15, 2018a: 13–21) where concept identity guarantees topic continuity. They argue topic sameness is got by people using the same concepts, which is in turn understood with people sharing a "tradition" for fixing the content of that particular concept. This tradition is supposed to include all the various representational and non-representational ways that people's thoughts are related to the world. Diverse Packages Theory is compatible with such a view because of the distinction it draws between a concept's invariable content and variable content, and because it is neutral on precisely how or which facts fix the invariable content of a concept. As such, the theory would say that if topic continuity depends

[18] See Sawyer (2020, 2018b,a), Schroeter and Schroeter (2014) and Cappelen (2018: Ch. 9) for an overview of this type of view of the relation between concepts and topics.

on concept identity, we should merely look at the invariable content of a concept, and identity there indicates that reasonable people "diverge" in their concepts in the sense of putting forth candidates for the same concept. Since Diverse Packages Theory is not committed to any particular story about how or which facts fix the invariable content of a concept it is entirely compatible with what Schroeter and Schroeter, and Sawyer say about how concept identity is determined.

Beyond views that cash out topic sameness through concept identity, Diverse Packages Theory is also compatible with a range of other views as well. For instance, it is entirely compatible with a view like Cappelen's (2018: 107–108) that addresses topic continuity with the idea that it is a pre-theoretic notion. This is because Diverse Packages Theory does not use topic continuity as a theoretical notion or commit to any broader metasemantic theory about entities other than, concepts, meaning, extension, and intension. As such, it is neutral on what pre-theoretic notions can be posited to make sense of topic continuity. This means it is compatible with the solution offered by Cappelen which makes use of the notion of "samesaying" as a pre-theoretic course-grained notion independent of a word's meaning or concept.

Finally, Diverse Packages Theory is also compatible with a purely pragmatic view of topic continuity of the sort argued for by Roberts (2012). For instance, the topic for a reasonable disagreement could be supplied by "questions under discussion". These are questions that set the function of the discourse between the parties. The initial question is then something like "What institutions and outcomes does justice require?". This would then elicit conflicting answers from reasonable people. This in turn gives rise to further questions under discussion that are ever more general to finally answer the initial question. On this view what seems like discontinuity of topic is really the work of pragmatic mechanisms in the disagreement to ultimately answer the initial question. Diverse Packages Theory is entirely compatible with a view like this because it has taken no position on the pragmatic mechanisms involved in a disagreement beyond those involved in metalinguistic negotiations. To that end, it is entirely compatible with additional pragmatic mechanisms if need be.

6 Conclusion

In this chapter I argued for and defended Diverse Packages Theory as the best explanation of reasonable disagreement about justice. I did this first by motivating the need to read cases of reasonable disagreement as either canonical disputes or as metalinguistic negotiations. I argued this allows us to read cases of reasonable disagreement that

are both ordinary and deep disagreements as genuine. I then showed how we can explain why reasonable disagreements that are either canonical disputes or metalinguistic negotiations occur by describing how people possess and use diverse 'concept-conception packages' to make moral and political judgements. I then showed how these moving parts yield an explanatory model that can explain both ordinary cases of reasonable disagreement and cases of reasonable disagreement that are deep disagreements.

I then argued that, aside from its ability to actually explain the cases it purports to explain, the theory has two comparative advantages. The first advantage is that by being able to explain deep disagreements, it is more explanatorily powerful than extant theories that claim to explain reasonable disagreement. The second advantage is that it relies on a more parsimonious theory of concepts than the only extant explanation of deep disagreements. I then considered and responded to three objections that might be put against Diverse Packages Theory. I argued that none of the objections warrant rejecting Diverse Packages Theory.

With Diverse Packages Theory now on hand as the best explanation of reasonable disagreement, in the next two chapters I put it to use in the second stage of my Disagreement to Legitimacy argument. I will evaluate whether political liberalism or political realism can, in light of Diverse Packages Theory's explanation, show how a society can achieve a stable political order. This is what I turn to in the next three chapters.

Chapter 4

The Instability of Political Liberalism

1 Introduction

At the end of the last chapter, I concluded that the best explanation of reasonable disagreement was Diverse Packages Theory. This meant that the best explanation of reasonable disagreement about justice – the sorts of disagreements that matter for Disagreement to Legitimacy arguments – was that reasonable people possess and use diverse concept-conception packages of justice which cause them to make conflicting judgements about the institutions and outcomes justice requires. With this explanation on hand, the rest of this thesis puts it to use in the second stage of my Disagreement to Legitimacy argument. This second stage is what I take up now.

The goal of this chapter and the next, is to argue that, in light of Diverse Packages Theory, extant conceptions of political liberalism and political realism cannot show how reasonable people can achieve a stable political order. This will serve as a negative argument for the Dual Convergent Conception I argue for in Chapter 6.

To get clear on how the argument will proceed, recall two points from Chapter 1. The first point is that the ability to show how reasonable people can achieve a stable political order is the metric of evaluation for a theory of political legitimacy, and therefore for evaluating conceptions of political liberalism and political realism. This is because the *inability* of reasonable people to achieve a stable political order in the face of reasonable disagreement about justice is the problem that theories of political legitimacy are solutions for. The second point to recall is that stability as a metric is understood along two dimensions: the ability to show how to create a political order and the ability to show how to maintain this political order over time.

The ability to create a political order amounts to showing how all reasonable people's balance of reasons can provide them sufficient moral reason to coordinate on

a political principle or rules that are coercively enforced despite their reasonable disagreements about justice. The ability to sustain a political order amounts to showing how all reasonable people's balance of reasons can be maintained so they continue to have sufficient moral reason to coordinate on conceptions of justice or rules that are coercively enforced despite the endogenous and exogenous forces disturbing their balance of reasons. The forces we are concerned with are limited to, in the case of the endogenous forces, the sphere of activity the political principles or rules themselves permit or encourage and, in the case of the exogenous forces, the normal changes in circumstances a political society is likely to face.

The argument over the next two chapters involves showing how the various extant conceptions of political liberalism and political realism fail on at least one of the two dimensions of stability. The goal of this chapter more specifically is to argue that the two main conceptions of political liberalism – the Consensus Conception and Convergence Conception – fail on at least one of the two dimensions of stability and that this motivates the general move towards political realist theories of legitimacy.

Recall from Chapter 1, the general strategy that unites conceptions of political liberalism is proposing that a political principle or rule is legitimate if all reasonable people conclusively justify *endorsing* it as one's own. This is the normative standard of 'public justification' or "Public Justification Principle" formalised by political liberals (Gaus and Vallier 2009). The normative standard has, broadly speaking, the following structure. The first part is that the object of justification is either a principle or rule. The second part is that the nature of conclusive justification is context insensitive. This is because the facts that justify a principle or rule are purely the content of reasonable people's moral reasons. The third part is that the attitude towards the political principles or rules that are justified is endorsement. This is to say that when a principle or rule is publicly justified it is internalised as a requirement of justice or morality broadly construed. As we shall see, political liberals then differ on whether the objects of justification are only general principles for the design of institutions or context and issue specific moral rules, and on the specific way conclusively justification is achieved. Nevertheless, on all conceptions of political liberalism, a stable political order is achieved by reasonable people's balance of reasons providing them and continuing to provide them sufficient moral reason to endorse a political principle or rule. But, as I'll argue, given Diverse Packages Theory's explanation of reasonable disagreement about justice, conceptions of political liberalism and its normative standard of public justification cannot show reasonable people how to achieve a stable political order.

To that end, the rest of this chapter proceeds as follows. In §2 I explain, the Consensus Conception political legitimacy, why it fails to show how reasonable people can create a political order, the responses available to political liberals and my reply to each of the responses. In §3, I explain the Convergence Conception of political legitimacy, why it fails to show how reasonable people can maintain a political order, the responses available to political liberals and my reply to each of the responses.

2 Consensus Conception

The most prominent version of political liberalism is the Consensus Conception. On this conception political liberalism's normative standard of public justification is understood to be targeted at political principles, rather than specific rules, and reached by a consensus of moral reasons. This notion of public justification then forms the heart of the Consensus Conception as the grounds of stability. The Consensus Conception proposes that understanding public justification as an overlapping consensus shows how reasonable people's balance of reasons can provide them sufficient moral reason to endorse and continue endorsing a political principle over time. In short, that it shows how reasonable people can achieve a stable political order. In this section I argue that the Consensus Conception cannot do this.

To that end, this section proceeds as follows. In §2.1 I lay out the Consensus Conception of political liberalism. In §2.2 I argue that the conception, in light of Diverse Packages Theory, faces a new sort of objection: the Conceptual Inconclusiveness Objection. In §§2.3–2.4 I consider two ways political liberals could respond to the objection. I argue that in both cases the responses fail or bring on their own problems. As a result, I conclude that the Consensus Conception of political liberalism cannot show how reasonable people can achieve a stable political order.

2.1 The Theory of Political Legitimacy

The Consensus Conception of political liberalism is best summarised as:

Consensus Conception: A political principle is legitimate if there is an overlapping consensus of reasons amongst all reasonable people that conclusively justifies endorsing it.

This is the theory of political legitimacy that Rawls (2005) and his followers argue for.[1] The core idea is that no matter what moral reasons reasonable people have for endors-

[1]For those who follow Rawls, see Nussbaum (2011), Hartley and Watson (2018, 2009), Larmore (1999, 1990), and Leland and van Wietmarschen (2017). See also, Lister (2013) although he calls the

ing their comprehensive conceptions of justice, or making judgements about what justice requires, there is a set of reasons they all share or "overlap" on, which conclusively justify political principles. These shared set of reasons that all reasonable people overlap on are such that they override all other justice related reasons and conclusively justify endorsing a specific family of political principles. These are political principles that are part of "political conceptions of justice". As such, political principles that are part of political conceptions of justice are ones that all reasonable people ought to obey and that are morally permissible to enforce with the use of coercive power. This is because, all reasonable people have sufficient moral reason to endorse any particular principle that is a member of the family of political conceptions of justice. This is ultimately what is captured by Rawls's (2005: 137) famous liberal principle of legitimacy: "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." The principles and ideals that constitute a conception of justice are legitimate according to Rawls if all reasonable people can endorse them. To that end, the Consensus Conception shows how reasonable people can achieve a stable political order by cashing out the core idea above with the idea of an overlapping consensus, and the educating role of a political conception of justice's principles and practical ideals.

As Rawls (2005: 144–149) argues, reasonable people can create a political order despite reasonable disagreement about justice because all reasonable people can find an "overlapping consensus" on a set of reasons that conclusively justify political principles that are part of a political conception of justice. The overlapping consensus justifies political conceptions because it involves a consensus on a particular subset of moral reasons – "political values" – that are reasons related to the domain of the political. This is the limited domain of social life in which people are related to one another within the structure of their society's basic social institutions, which they cannot enter and exit easily or transactionally, and in which the power of those institutions is always coercive power. Outside this limited domain, there will be a host of moral values that will serve as reasons for how people ought to behave in their lives and over which reasonable people will disagree. But, for issues in the domain of the political the theory says there is a limited set of political values that all reasonable people share.

As Rawls (2005: 64, 140) explains, the consensus on these political values conclu-

Consensus Conception the "reasons-for-decisions model", and Quong (2011); although his "alternative view" differs in structure from Rawls's original account it is still committed to the underlying consensus of reasons that is distinctive of the Consensus Conception.

sively justify political conceptions of justice in two stages of justification. The first stage is a "freestanding" pro tanto justification of the political conception. What this means is that the political values are constructed out of fundamental ideas that are implicit in the shared political culture of the society in which reasonable people live rather than out of any comprehensive conception of justice. The public culture consists of the existing political institutions and the historical texts and traditions of their interpretations of a democratic constitutional society. The ideas taken from this culture are the idea of society as a fair system of social cooperation (with its criterion of reciprocity), the idea of people as free and equal (with their capacity for a conception of the good and sense of justice), and the idea of a well-ordered society (marked by the fact of reasonable pluralism and regulated by a conception of justice all endorse). This pool of shared ideas provides the grounds for constructing the political values in a way that is "freestanding" from reasonable people's comprehensive private conceptions (or as Rawlsians call them "reasonable comprehensive doctrines"). The political values justify a political conception of justice in a way that is separate from and disconnected from the non-political values of reasonable people's comprehensive conceptions of justice.

These political values *conclusively* justify a political conception of justice because reasonable people can embed it within their own comprehensive conceptions of justice and see that it is still justified from within that larger set of moral values. This is the second stage of justification, which occurs for two reasons. The first is that the political values are very weighty values that override all other values when it comes to issues relating to the use of coercive political power. They do this because, as Rawls (2005: 139) says, they are related to the "the basic framework of social life" and the "very groundwork of our existence". They are values that specify "the fundamental terms of political and social cooperation" necessary for pursuing all our other values and considerations. This shows how reasonable people's balance of reasons gives them sufficient moral reason to endorse a political conception of justice and the political principles that constitute it. A conception of justice justified by the political values is tailored for a specific purpose, namely to settle the constitutional essentials and matters of basic justice in a society. For that purpose the political values are overriding.

The second reason is that in virtue of the freestanding justification – the construction of the political values from ideas implicit in the shared political culture – reasonable people can see the only way to use coercive power in a way that satisfies the criterion of reciprocity is to endorse and defer to the political conception. As Rawls (2005: xliv–xlvii, 50) explains, the criterion of reciprocity is that the "exercise of polit-

ical power is proper only when we sincerely believe that the reasons we offer for our political action may reasonably be accepted by other citizens as a justification of those actions". Since the criterion of reciprocity comes from ideas in the shared political culture, reasonable people are already committed to it when they take into account other reasonable people. Therefore, it is only the political values constructed in a free-standing way that can justify coercively enforced political principles when there are reasonable people with their own comprehensive conceptions. With all that a political principle is conclusively justified when all reasonable people can come to see it as justified from within their own comprehensive conceptions of justice. To that end it can create a political order despite reasonable disagreement about justice.

According to the theory, this political order can be maintained by the educating role of the political principles and the practical ideal of public reason that the political values justify. As Rawls (2005: 81–86) argues the political principles that are part of a political conception of justice design social institutions that help reasonable people maintain their sense of justice and hence enable them to maintain their balance of reasons to endorse the political conception of justice. This means the political principles educate citizens to maintain their sense of justice which allows them to maintain their effective endorsement of a political conception of justice. The political principles can play this role because, as Rawls (2005: 83) claims, reasonable people have a "reasonable moral psychology" which means they have a desire to fully realise their capacity for a sense of justice and capacity for a conception of the good.[2] This means they also have a capacity to acquire conceptions of justice and act as these conceptions require. As a result they are receptive to the effective regulation of a society by the principles of a political conception of justice.

The second way the political order is maintained is that the political values that reasonable people are supposed to overlap on also justify a practical ideal of public reason. As Rawls (2005: 215–227) argues, the idea is that the political values also justify an ideal of democratic citizenship. This ideal of citizenship is that reasonable people ought to discuss and interact within the institutions of the political conception of justice by using coercive power only with reference to the political values. This means that reasonable people's balance of reasons in favour of the political conception of justice is maintained by reasonable people taking seriously the fact they disagree with other reasonable people and given their sincere desire to agree (which is part of their reasonable moral psychology as reasonable people) they see that the way to reconcile

[2]Note this moral psychology is a philosophical view of what reasonable people are motivated to do and develop. It is not an empirical view.

with each other so that all have sufficient reason to coordinate on the use of political power is to decide political matters according to the political values that they all share. With those two elements – the educating role of the political conception's principles and its practical ideal of public reason – the theory shows how reasonable people can sustain a political order.

All in all, political legitimacy on the Consensus Conception of political liberalism involves cashing out how public justification can be achieved with the idea of an overlapping consensus, the educating role of a political conception's principles and the practical ideal of public reason. Reasonable people can achieve a stable political order because those ideas show how reasonable people's balance of reasons can provide and continue to provide all of them with sufficient moral reason to endorse political principles despite their reasonable disagreements about justice.

2.2 The Conceptual Inconclusiveness Objection

The problem with the Consensus Conception is that Diverse Packages Theory shows reasonable people can genuinely disagree about justice without having any overlapping consensus of reasons. This is because reasonable disagreements about justice can be caused by reasonable people possessing and using divergent concepts of JUSTICE. Possessing divergent concepts of JUSTICE means that people will have diverse considerations when deliberating and forming conceptions of justice. This means the set of "political values" are not going to necessarily be shared by reasonable people. In fact the set of political values will vary for reasonable people with different concepts of justice. As a result there would be no unique set of values that conclusively justify a single conception of justice for reasonable people. Which means the theory would legitimate multiple incompatible political principles. I call this the Conceptual Inconclusiveness Objection. It is a version of the inconclusiveness objection, already made by some theorists, which operates at the level of concepts.[3] This means that the Consensus Conception cannot show reasonable people how to create a political order because it cannot show how their balance of reasons can provide them all sufficient reason to endorse the same political principles.

The Conceptual Inconclusiveness Objection is significant because it is a much stronger version than the sort pressed by theorists in the literature so far. For instance, Reidy (2007: 261, 2000: 63–70) has argued that the set of political deliberative con-

[3]See Boettcher (2015: 194–195) for a good discussion of how the inconclusiveness objection, being a variety of the broader incompleteness objection, is similar, but importantly different to the objection put by others in the literature.

siderations that are supposed to be the subject of the overlapping consensus do not contain any conclusive reason for the use of coercive power in some cases like abortion or animal rights. But, the version of the objection I present here is stronger than Reidy's. For Reidy the problem is that it is *unlikely* there is any rational way to determinately weigh the shared reasons and that there are political issues the resolution of which depends on the resolution of other background issues.[4]

My objection however is that Diverse Packages Theory shows reasonable disagreement about justice can be produced by differences in the very conceptual ingredients that people use to deliberate about justice. This is because what deliberative considerations people share depends on the concept of JUSTICE they possess and use. Parties that possess and use divergent concepts will have sets of conflicting considerations inputted into their deliberations about endorsing conceptions of justice. For instance, reasonable people who are anti-democrats, or in favour of more limited forms of democracy will not share the "political values" constructed out of the ideas implicit in their society's political culture. The point is that given Diverse Packages Theory we cannot, as the Consensus Conception contends, assume that reasonable people will share a set of deliberative considerations that override all other considerations.

The consequence of the Conceptual Inconclusiveness Objection is that it shows how on the Consensus Conception a society will be disordered in one of two ways. On the one hand, when reasonable people do not share a concept of JUSTICE they split into groups where each group endorses and acts in accordance with the political principles that are conclusively justified to them. As a result the society falls into anarchic disorder because they oppress each other by imposing the political principles that are justified to each of them. On the other hand, reasonable people might refrain from oppressing those with incompatible concepts, but have unresolved claim disputes which create disorder when individuals try to pursue their conceptions of the good life. This means the theory cannot show us how reasonable people can achieve a stable political order.

2.3 Broadening the Scope Response

One way political liberals might respond to the Conceptual Inconclusiveness Objection is to broaden the scope of public justification. The idea is that the normative standard of public justification should be applied to reasonable people's concepts as well such that legitimacy is grounded in a consensus at a deeper level of thought. This

[4] See Williams (2000) for discussion of Reidy's version of the objection and to see how it differs from the version I present here.

sort of response is what it seems political liberals who argue for a general "accessibility" constraint on reasons are aiming for (Boettcher 2015: 200–205; Eberle 2002: 253–359). On this view of the overlapping consensus, the content of the consensus is comprised of the reasons that all reasonable people can access with their cognitive faculties even if the deliberative considerations they actually access are all mutually incompatible. The idea is that when the object of consensus is reasonable people's conceptual content we can make sense of a greater range of admissible deliberative considerations and therefore make the Consensus Conception consistent with Diverse Packages Theory. But, this response does not work because given Diverse Packages Theory's explanation of reasonable disagreement about justice, there are at least two ways to understand what it means to apply public justification to reasonable people's conceptual content and both fail.

One way to apply public justification to conceptual content is to say that although reasonable people can possess divergent concepts, they ought to only use the concepts that sufficiently overlap in content with the concepts that other reasonable people use. There are two problems with this. The first problem is that it seems prima facie implausible that all of our reasonable disagreements will involve overlaps in conceptual content. It might work for modest conceptual differences where reasonable people only differ slightly in their conceptual content such that they disagree over certain hard or borderline cases. But, there is no reason to think that the forms of deep disagreement that Diverse Packages Theory explained in Chapter 3 will be like this. For instance, it might work when people disagree about precisely at which point a fetus counts as a moral person. To use only the conceptual content they overlap on may mean only using the conceptual content that relates to the general window in which fetuses become moral persons. But, there is no reason to assume that all our reasonable disagreements will be like this. For instance, as I describe in Barry and Nora's disagreement in Chapter 3, deep disagreements involve cases where reasonable people reject the other's deliberative consideration entirely and therefore the conceptual content that gave rise to it. In short, reasonable people can disagree in such a way that they do not overlap in conceptual content at all.

The second problem is that although searching for deeper conceptual overlap is plausible, it is implausible to ask of people if the conceptual content in question is useful for making judgements that are unrelated to justice. For instance, two reasonable people might disagree about justice, not by possessing divergent concepts of JUSTICE, but by possessing divergent concepts of MORAL PERSONHOOD. This would be a case where they broadly overlap on the conceptual content that moral personhood is a

relevant consideration for deliberating about justice, but do not share the particular content of that consideration that specifies what counts as a moral person. In these cases they might have good reason to use the concept of MORAL PERSONHOOD because of its usefulness for other circumstances. Perhaps in circumstances where they are advocating for their comprehensive conception of justice or the good life. In these cases it seems implausible to expect them to not use the conceptual content they have very good reason to use.

However, another way to apply public justification to conceptual content might be to say that public justification applies to the *formation* of our concepts themselves and hence to tacit beliefs about the facts that fix the conceptual content of our concepts. On this view, public justification requires reasonable people share these tacit beliefs. This will allow reasonable people to then possess the same concepts.

There are two problems with this. The first problem is that it is implausibly demanding. Our beliefs about the facts that fix the content of concepts are perhaps even more diverse than our conceptual content and so finding consensus there is unlikely. This is because an important part of Diverse Packages Theory's explanation of reasonable disagreements was that differences in conceptual content were explained by differences in people's intrinsic dispositions and personal experiences towards certain facts as the facts which fix the content of their concepts. This final cause is not something that can plausibly be bracketed or be the object of consensus because it is a plain empirical fact that people are dispositionally diverse and have very diverse personal experiences. The strategy of constantly moving back the overlapping consensus to deeper levels of thought is not likely to present less diversity, but likely more of it.

The second problem is that it also seems self-undermining for a *liberal* conception of legitimacy since it would be committed to a profoundly illiberal view in the topic of Conceptual Ethics.[5] This is the area of philosophy that deals with the ethics of possessing and using certain concepts over others. A political liberal taking up the response above would be committed to denying reasonable people the freedom of conscience and thought. This means it requires an illiberal standard on people's concept formation in order to legitimate political principles that are part of liberal political conceptions of justice. In sum, the strategy of broadening the scope of public justification is an inadequate response to the Conceptual Inconclusiveness Objection.

[5] See Burgess and Plunkett (2013a,b) for an overview.

2.4 Narrowing the Constituency Response

Another response to the Conceptual Inconclusiveness Objection might be the sort advanced by Johnathan Quong and Micah Schwartzman where the idea is to narrow the constituency of people that count as reasonable.[6] On this strategy, Diverse Packages Theory's explanation of reasonable disagreement about justice is neutralised at the level of *who* counts as reasonable and not, as the Broadening the Scope Response tries, at the level of what is publicly justified.

The response starts with the idea that reasonable disagreement about justice is, by definition, a special form of disagreement. For Quong, reasonable disagreements about justice are justificatory disagreements and not foundational ones. As Quong (2011: 214) says:

> Reasonable disagreements about justice are thus justificatory by definition. The truth of this claim does not rest on any empirical claim about substantive agreement between actual citizens on principles of justice at any level of abstraction. Rather, reasonable disagreements about justice are justificatory by definition because they must always involve reasonable citizens who share a commitment to the public or political values that are the subject of the overlapping consensus...

In contrast, foundational disagreements are those where people do not share a commitment to the public and political values. Quong (2011: 206) thinks disagreements about the good life will fall into this category whilst reasonable disagreements about justice will not.

For Schwartzman (2004: 200–201) reasonable disagreements about justice are disagreements about issues nested within the broad notion of a public political conception of justice, and not about any issues outside that broad notion. This is because Schwartzman argues given the definition of a well-ordered society as one where "everyone accepts and knows that others accept the same principles of justice" political principles outside that broad notion cannot order such a society. As Schwartzman (2004: 200) says:

> Principles of justice that cannot be interpreted in ways consistent with liberal freedoms are inadmissible because they cannot serve as the basis of a well-ordered society. They do not appear in the nested set of public political conceptions about which citizens may reasonably disagree.

[6] Strictly speaking, Quong's target is the Asymmetry Objection, but he employs the same sort of response to inconclusiveness objections.

This means that reasonable disagreement about justice is by definition limited to disagreements about "which principles of justice are the most reasonable, and…whether the most reasonable principles have been satisfied" and not about what states of affairs are just or which principles are the true principles of justice.

In both cases the idea is that in virtue of being a disagreement of a certain kind reasonable disagreements about justice involve people who share a standard of justification for their judgements. In the case of reasonable disagreements about justice people will, in virtue of being reasonable, share and use public justification as the standard for making their judgements. This means that the way Diverse Packages Theory's explanation gives rise to the Conceptual Inconclusiveness Objection is largely irrelevant. This is because the public justification of a political conception of justice is only inconclusive for unreasonable people and not for reasonable people.

But this response does not work because applying the distinction between foundational and justificatory disagreements that Quong and Schwartzman assert to vindicate the Consensus Conception opens it up to the Public Dogma Objection (Campos 1994; Macedo 1991; Besch 2012).[7] This is because it shifts the goal posts on what it means to be reasonable without any independent reason. It turns public justification from a normative standard that reasonable people adopt to achieve a stable political order into a constitutive feature of reasonable people themselves. This means that the argument for the Consensus Conception rests on the prior acceptance of public justification and therefore by those it applies to. This would make the Consensus Conception a form of "public dogma" or "secular fundamentalism" (Besch 2012: 165; Campos 1994: 1824; Macedo 1991: 58). It excludes those that have not yet endorsed public justification from the constituency to which public justification is supposed to apply. Its normative standard legitimates the use of coercive political power against putatively reasonable people who are necessarily excluded from having that power justified to them. This means that the way the Consensus Conception achieves a stable political order is trivial. The balance of reasons it aims to create will always be achieved since it will only be concerned with people who already endorse public justification. And further, it will always maintain a political order because as soon as reasonable people question or try to reevaluate the justificatory grounds of public justification

[7] Schoelandt (2015: 1037–1041) makes a similar point, but insists the problem is that restricting who counts as reasonable means political liberals have to reject the interpersonal type of justification distinctive of political liberalism in favour of an impersonal type of justification. I do not think we have to go that far. Rather the problem is that, in Shoelandt's terms, the response tries to restrict an interpersonal type of justification to the extent it generates its own instability. As such it is not suited as a theory of political legitimacy.

they will drop out of its constituency and so the balance of reasons is maintained by default.

Before moving on, let me dispel a reply that theorists like Quong (2011: 140), who hold the "internal conception" of political liberalism, may be tempted to make.[8] The internal conception of political liberalism is that political liberalism as a theory of political legitimacy should be understood as justifying liberal political principles to liberals themselves. As such, it is a project that is internal to liberal political theory. Now, those like Quong who profess this view of political liberalism may be tempted to say that the distinction between justificatory and foundational disagreements does not assume reasonable people accept anything as demanding as public justification, but rather only some "basic liberal norms or values", like "the moral ideal of persons as free and equal, and of society as a fair system of cooperation".[9]

But of course this does not avoid the objection because "the moral ideal of persons as free and equal, and of society as a fair system of cooperation" are fundamental ideas in a public culture that both liberals and non-liberals can share.[10] In fact this is crucial to the Consensus Conception. Although liberals will in addition also share some "basic liberal norms or values", non-liberals will not. All reasonable people in such a society will claim to be committed to the view that people are free and equal and to fair terms of social cooperation. They may not share the specific liberal interpretation of these values, but that is only because they are after all non-liberals and so do not possess comprehensive liberal conceptions of justice. To that end, the Narrowing the Scope response also fails to rebut the Conceptual Inconclusiveness Objection.

2.5 Conclusion

I have argued in this section that given Diverse Packages Theory's explanation of reasonable disagreement about justice, the Consensus Conception of political liberalism cannot show reasonable people can achieve a stable political order. I argued that it faces the Conceptual Inconclusiveness Objection which means it either legitimates multiple political principles or does not justify any of them. In both cases, it cannot show how there can be a balance of reasons amongst reasonable people that provides

[8] Andrew Lister (2013) also seems to endorse the internal conception, but in a communitarian version. See also Schwartzman and Wilson (2019) on their use of what they term "liberal reasonableness".

[9] See Nussbaum (2011), Wenar (1995) and Kelly and McPherson (2001) for a similar point. See also Leland and van Wietmarschen (2012) for whom the acceptance of basic liberal values is necessary, but not sufficient for reasonableness.

[10] See Fowler and Stemplowska (2015) and for a similar point that is originally made in defence of the Asymmetry Objection.

them sufficient moral reason to endorse the same political principles.

I then argued that broadening the scope of public justification fails as a response because it is unclear that public justification can be applied to people's conceptual content. I then argued that the narrowing the constituency of public justification fails because there is no independent reason for the distinction required for doing so and as a result makes achieving a stable political order trivial. Therefore, both of the ways that political liberals can try to avoid the Conceptual Inconclusiveness Objection do not work and so the Consensus Conception cannot provide an adequate theory of political legitimacy.

3 Convergence Conception

Of course the Consensus Conception is not the only way to understand political liberalism's normative standard of public justification. Another way to understand it is through the idea of a 'convergence' of reasons.[11] The core idea is that 'public justification' is achieved through all reasonable people having *some* reason that conclusively justifies issue and context specific moral rules. This is the idea of public justification that forms the heart of the Convergence Conception. It proposes that the convergence of reasonable people's conclusive justifications shows how their balance of reasons can provide them sufficient moral reason to endorse issue and context specific moral rules and continue endorsing these moral rules over time. In short, it shows how reasonable people can achieve a stable political order.

By allowing far more reasons to count as conclusive and focusing on issue and context specific moral rules as opposed to general political principles, the Convergence Conception can provide a way for political liberalism to avoid the Conceptual Inconclusiveness Objection and the series of problems that arise from responding to it. But, as I'll show, given Diverse Packages Theory, the Convergence Conception cannot show how reasonable people can achieve a stable political order. The type of agreement that is supposed to represent reasonable people's balance of reasons providing them sufficient moral reason to endorse moral rules cannot be maintained. This means, the conception cannot show how reasonable people can maintain a political order.

To that end, the rest of this section proceeds as follows. In §3.1 I detail the Convergence Conception and how it can avoid the Conceptual Inconclusiveness Objection.

[11] See Thomas Nagel (1987) and Fred D'Agostino (1996) for the earliest explication of the distinction between the Consensus and Convergence Conceptions of political liberalism.

In §3.2 I argue the Convergence Conception faces what I call the Verbal Agreement Objection. In §3.3 I consider two ways that political liberals could respond to this objection by saying more about the precise mechanism of convergence. In §3.3.1 I detail the Multi-Perspectival View and argue that it is not an adequate version of the response because it encourages people to back out of agreements when it no longer favours them, requires complex calculations of the effect of a bargain, and requires a particular conception of individual rights that will itself be the subject of reasonable disagreement. In §3.3.2 I detail the Social Equilibrium view and argue it is also not an adequate version of the response because it faces the Conceptual Integrity Objection and so generates its own sources of instability in maintaining a political order. As a result I conclude that although the Social Equilibrium view is the best version of the Convergence Conception, political liberals who take it up are caught between the instability of the Verbal Agreement Objection and the Conceptual Integrity Objection. This gives us good reason to reject political liberalism and consider other types of theorises of political legitimacy, like political realism.

3.1 The Theory of Political Legitimacy

The Convergence Conception of political liberalism can be summarised as:

Convergence Conception: A set of issue and context specific moral rules are legitimate if there is a convergence of mutually intelligible reasons amongst all the reasonable people in a society that conclusively justifies endorsing them.

This is the theory of political legitimacy that theorists like Gerald Gaus (2016, 2011b)[12], Kevin Vallier (2019, 2014, 2011; 2009) and arguably Ryan Muldoon (2016) have argued for.[13] Although they all differ on the specific mechanism that underlies the political order the conception is supposed to help us achieve (I will return to this point in §3.3), the core idea is the same. That is, the public justification of issue and context specific

[12] Strictly speaking Gaus's theory concerns social morality and the construction of a moral order, with a political order being a tool for constructing and maintaining the moral order (Gaus 2016: 177–187, 206–207, 2011b: 460–470, 545–546).

[13] As Gaus (2016: 168) notes, Muldoon sees his view as an alternative to political liberalism, but that is because he only considers the Consensus Conception of political liberalism. Arguably, Amartya Sen (2010) also advances a sort of Convergence Conception. But, as Gaus (2016: 155–163) convincingly argues, Sen's view sits somewhere between the Consensus and Convergence Conceptions of political liberalism. Sen is still concerned with conceptions of justice and is still committed to some consensus in how reasonable people view their shared social world. Given that, I leave aside discussing Sen's view directly even though we should keep the view in mind as a marker of how theories of legitimacy exist on a continuum between consensus and convergence.

moral rules that are coercively enforced is a matter of all reasonable people converging in their conclusive justifications through potentially different and unique reasons.

Now, there are two important ways that the Convergence Conception is different from the Consensus Conception. Both of which combine to help political liberals avoid the Conceptual Inconclusiveness Objection. The first way is that public justification is cashed out as the convergence of mutually intelligible reasons. As Gaus and Vallier (2009: 58–59) summarise when contrasting consensus and convergence:

> Contrast this to the Convergence Conception according to which members of the public may arrive at common laws by reasoning based on diverse values and concerns. Here pluralistic reasoning is the very basis of justification. As long as intelligibility obtains, all members of the public acknowledge that everyone engages in genuine reasoning such that each person's conclusions provide her or him with reasons to accept the law. So everyone can see everyone else as a self-legislator and freely subject to the law. Appealing to a law justified in this manner respects each person as free and equal, without any insistence that we reason in the same way.

Although Gaus and Vallier are using their own terminology for reasonable people – "members of the public" – the basic point is the same. When reasonable people can recognise the reasons others provide for endorsing a rule *as reasons* for endorsing it even if they do not see them as *good* reasons or conclusive reasons for themselves, then such reasons are intelligible. From there, the use of coercive power is publicly justified if all reasonable people can find *some* conclusive reason, that is within that pool of intelligible reasons, to endorse it. This fundamentally breaks from the Consensus Conception of political liberalism by rejecting the need for a shared set of reasons, and only requiring a convergence of mutually intelligible reasons.

The second way the Convergence Conception is different from the Consensus Conception is that the object of justification is not a political principle, but rules that concern specific issues in specific contexts which all can recognise as *moral rules*. Which means recognising them as requirements of justice or morality broadly construed. These moral rules can then play all sorts of roles from being constitutional laws to specific laws of the property-rights system. But, they must be issue and context specific. They are not general principles for how to construct rules or social institutions. Rather they are the moral rules that comprise social institutions themselves. This narrowing of the focus of public justification on issue and context specific moral rules is for two reasons. The first reason is that convergence on entire conceptions of justice is taken to be too unlikely given that reasonable people's powers of reason are

limited. As Gaus and Vallier say:

> If we are to have a good grasp of the public justifiability of moral demands based on moral rule L, we must have comparative knowledge of how L stacks up against the alternatives we can canvass. Even Members of the Public, who recognize their sufficient reasons, are of limited rationality, and as creatures of limited reasoning powers, when asked whether they have sufficient reason to endorse a rule, they must ask, "what are the alternatives we are deciding between, and what are the costs of refusing to endorse any of them?"...To ask "what do I have sufficient reason to endorse?" when I do not know the set of options is an ill-formed question. In the terms of economics, one must know the opportunity costs of one's choice – the value of the options that one has passed over – before one can come to any reasonable judgment of what is to be endorsed. (Gaus 2011b: 269)

> Our aim in constructing a justified lawmaking systems is, as far as possible, for acts of legislation to reflect what citizens understand as distinct and manageable issues. Here, as elsewhere, holistic justification is outside the bounds of real human reason. (Gaus 2011b: 496)

> We [convergence theorists] focus on the public justification of moral rules because they're the kind of social practice that can be internalized by most moral agents. Moral life is not based on generic moral principles like Rawls's difference principle, but on local rules governing local behavior. (Vallier 2019: 175)

The point here is that when people are limited in their powers of reasoning they cannot, given their differences in their concepts of justice, adequately decide whether they have sufficient reason to endorse a general political principle against all the other alternatives. Doing this with context and issue specific moral rules is easier and so convergence is more likely.

The second reason for narrowing the focus of public justification is that issue and context specific moral rules are what's needed for moral rules to do the work of effectively ordering a society where people who disagree cooperate and do not cheat. As Gaus (2011b: 113) says:

> ...group cooperation requires norms or rules that are specific enough in their requirements that cheater detection is highly reliable within the

> group. Cooperative rules must be sufficiently general so that they provide guidance in unforeseen future circumstances while, at the same time, it is reasonably clear to the great majority of group members (i) when the rule applies and (ii) what the rule calls for. Because punishment seems so fundamental to the evolution of cooperative orders, and because these two desiderata are so important to the emergence of effective punishment, it appears less important that rules be fine-tuned to get the "correct cooperative result" than that these desiderata be fulfilled.

The point here is that rules that do not clearly state what behaviour they require and in the context in which they require them cannot help us detect those that do not cooperate. This would make it difficult for people to actually maintain sufficient reason to endorse moral rules and act according to them.

The point in all of this is that the objects of public justification are moral rules that apply in specific contexts about specific issues like coercive laws whether they be constitutions or ordinary statue law. They are not general principles encompassing many issues and meant to be applied to various contexts like Rawlsian or Utilitarian principles. Whether general normative principles of justice can be abstracted out of the set of issue and context specific moral rules is of course another matter. But, on the Convergence Conception whatever these general principle are they are not the objects of public justification.

To that end, the theory shows how reasonable people can achieve a stable political order with the idea of a convergence of potentially conflicting reasons on a set of issue and context specific moral rules. According to the theory reasonable people can create a political order no matter what concepts and conceptions of justice they have, since all that's required to publicly justify issue and context specific moral rules is reasonable people to have *some* mutually intelligible conclusive reason to endorse it even if they all disagree about what that reason is. The idea is that when the normalisation on reasons is lowered to the level of 'mutual intelligibility' far more possible reasons can conclusively justify including the one's provided by people's private comprehensive conceptions of justice. This shows how reasonable people's balance of reasons can provide sufficient moral reason for all to endorse issue and context specific moral rules despite their profound disagreements about justice.

All this provides a way for political liberals to avoid the Conceptual Inconclusiveness Objection.[14] If public justification is construed as each reasonable person conclu-

[14] Although convergence political liberals have not countenanced Conceptual Inconclusiveness Objection per se, they have noted related problems and how the Convergence Conception can overcome

sively justifying issue and context specific moral rules through their own potentially unique reasons, rather than through a shared set of reasons, then people possessing and using diverse concepts is not a problem. When reasonable people possess and use diverse concepts of justice they will have incompatible sets of deliberative considerations when deciding whether to endorse a moral rule. But, these incompatible sets of deliberative considerations are not a problem since conclusively justifying an issue and context specific moral rule does not require sharing reasons. Rather, as long as each reasonable person finds some deliberative consideration to be a conclusive reason to endorse a moral rule, that moral rule will be publicly justified. Therefore, the Convergence Conception shows how despite reasonable disagreement about justice, reasonable people can create a political order.

According to the theory, this political order can be maintained because the convergence is not over any specific set of reasons. This means that when reasonable people's set of moral reasons changes or their weighing of them changes, this will simply change the set of issue and context specific moral rules that are converged upon. This means there is no point at which the publicly justified set of rules are oppressive or justify anarchic rebellion since whenever they become so, they will no longer be publicly justified since at least one reasonable person will not have conclusive reason to endorse them. Since the Convergence Conception holds on to the core idea that only publicly justified moral rules can be coercively coordinated upon, the political order is always maintained.

3.2 Verbal Agreement Objection

Whilst the Convergence Conception does well at avoiding the Conceptual Inconclusiveness Objection that plagued the Consensus Conception, some political liberals in favour of the Consensus Conception have pointed out that the way it avoids the objection is problematic. As Hartley and Watson (2018: 59–61) say:

> One way of understanding the convergence model of public reason is to read it as insisting on a kind of overlapping consensus for each particular law and not for a shared political conception of justice. So it appears an overlapping "convergence" on this law and that law, and so on, is a

them. See Gaus (2011b: Ch. 16) on the "Problem of Indeterminacy" and Vallier (2019: 114–115, 2014: 215) on the "Anarchy Objection". Interestingly, Boettcher (2015: 201–204) demurs and argues some form of inconclusiveness remains a problem because of the controversial notion of coercion assumed by convergence theorists. I leave aside Boettcher's objection and side with Gaus's (2014) response to similar objections as adequate.

> happy coincidence. But, the basis for social unity and continued commitment to seeking fair terms of social cooperation, is absent on this account. The contingencies of convergence – due to the "possibly fluctuating circumstances" connected with revisions of persons' views and, as a result, the laws that can be supported and the balance of power within a society – provide no deep assurance to a sustained commitment, by anyone, to laws that happen to find convergence at a given time.

Hartley and Watson's point is that the nature of convergent agreements and the objects of those agreements means that the political order the Convergence Conception shows how to create is chancy or coincidental. Even though it is not entirely unlikely given people's particular beliefs (ie. they could not have converged on anything else), it is chancy or coincidental that they converged on a particular moral rule because they could easily have converged on a different one with slightly different contexts or beliefs. This initial chanciness of convergence means it lacks a kind of normative stability. This is because, on the Convergence Conception, a political order is created by reasonable people agreeing to the issue and context specific moral rules rather than entire conceptions of justice. This means that when people's beliefs change or when the context changes the agreement will break down.

But, why should convergence theorists be particularly worried by this? After all, it is not new that people change their beliefs or that circumstances change. That is the nature of social and political life. Whilst this is true, I think there is something more to Hartley and Watson argument that should worry convergence theorists, namely that convergent agreements are *disposed* to break down, rather than being merely chancy. This is made clear when we look at what Diverse Packages Theory – the explanation of reasonable disagreement I developed in Chapter 3 – implies about the nature of convergence agreements. The basic idea is that an upshot of Diverse Packages Theory's explanation of reasonable disagreements is that at least some of the agreements that constitute 'a convergence on issue and context specific moral rules' are 'mere verbal agreements'. These are agreements grounded in the differences between the concepts that people possess and use. This is because, according to Diverse Package Theory concepts provide the deliberative considerations that people use to evaluate and endorse moral rules.

A convergence on an issue and context specific moral rule is then an agreement that some rule is a genuine moral rule on the basis of diverse considerations between reasonable people in a specific context. However, this means these agreements are highly sensitive to context change. This is because their agreement is grounded in the

convergence of different considerations and the specific contexts in which those considerations justify endorsing the moral rule. Slight context changes will mean that at least one person's considerations will not recognise the issue and context specific moral rule in question as a moral rule any longer. It will not be justified as a moral rule to endorse. The result of this is that at least some of the agreements that constitute a political order on the Convergence Conception are *disposed* to break down. This means, on the Convergence Conception, aspects of the political order that require endurance or quick institution building are by the very nature of their legitimacy likely to become illegitimate. I call this the Verbal Agreement Objection.

To get a clearer idea of what the objection is we need to understand what Diverse Packages Theory as the best explanation of reasonable disagreements implies about *agreements* and what causes them. Recall, from Chapter 3, Diverse Packages Theory is comprised of two core ideas. The first idea is that to vindicate reasonable disagreements as genuine disagreements they have to sometimes be read as canonical disputes, and sometimes be read as metalinguistic negotiations. The second idea is that causally explains reasonable disagreements read in either way is that reasonable people possess and use diverse concept-conception packages.

That analysis and explanation of reasonable disagreement starts with the premise that reasonable people's disagreements involve some practical conflict. It then explains this by citing how reasonable people's concept-conception packages diverge sufficiently to cause them to make moral and political judgements that are practically incompatible in the context in which they live.

But, Diverse Packages Theory could also explain cases where we start with the opposite premise. It could start with the premise that reasonable people make practically *compatible* moral and political judgements, and then explain this agreement by citing how reasonable people's concept-conception packages diverge in a convenient way such that they lead them to agree in the context in which they live.

Now, these sorts of agreements are nothing but the idea of convergence at the heart of the Convergence Conception where all reasonable people are said to endorse a moral rule for their own potentially unique conclusive reasons. Two parties endorse a moral rule despite disagreeing about what justice requires because, they each agree that the moral rule in question furthers their own particular view about what justice requires. This means that on the Convergence Conception, public justification ends up being a series of agreements where reasonable people are lucky enough that their concept-conception packages diverge in a convenient way such that they lead them to make moral and political judgements that are practically *compatible* in the context in

which they live.

So far so good. But, all this implies two ways that *agreements* between reasonable people are caused. The first way is that reasonable people agree by possessing and using concept-conception packages that diverge only in conceptions. They then agree in virtue of the differences in their conceptions of justice being sufficiently convenient that they entail endorsing the same issue and context specific moral rules.

Agreements like these is what I take Hartley and Watson to be worried about. They will of course be unstable to some degree, but convergence theorists have good reason not to be too worried. After all, in these cases reasonable people share concepts. As such they share a categorisation of their social world and what is a morally relevant consideration when agreeing to moral rules. This ensures that slight changes in context do not undermine the justification of a rule.

The problem for the Convergence Conception lies in the second way that Diverse Packages Theory explains agreements, namely that reasonable people can agree by possessing and using concept-conception packages that diverge in their concepts. This means that sometimes when reasonable people agree in the way the Convergence Conception proposes, they could agree wholly *in virtue of* the content of their particular concepts. This would be a case of a 'mere verbal agreement'.[15]

For example, Alice and Beth could agree to an issue and context specific moral rule that says resources ought to be distributed according to talents in their society. Alice believes the rule is what justice requires because it helps individuals and society flourish and, individuals and society flourishing is the only relevant consideration for a rule about distributing resources. On the other hand, Beth believes the rule is what justice requires because she believes it gives people what they morally deserve and moral desert is the only relevant moral consideration for a rule about a distributing resources. In such a case Alice and Beth's agreement would be a merely verbal agreement because they agree on the rule wholly in virtue of the particular concepts they possess and use. It is only in virtue of the distinct and unique considerations their concepts determine as morally relevant for deliberating about distributive rules that enables them to agree.

The problem with a merely verbal agreement like Alice and Beth's is that given they possess and use divergent concepts of JUSTICE, their agreement on an issue and context specific moral rule is highly sensitive to changes in context. It is highly sensitive to changes in the facts that constitute the moral rule, ie. changes in what a rule

[15] See Chalmers (2011: 525–526) for this way of construing verbal agreements. See also Ballantyne (2016) on construing verbal agreements through the same answer satisfying two different questions.

permits or prohibits as a matter of fact. For instance, a slight change in what counts as a talent, suppose Beth discovers people have a natural talent to cheat others, might cause the agreement to break down because cheating others does not promote moral desert. As such Beth would not see the rule as conclusively justified anymore given what she views as morally relevant in her social world, namely moral desert and nothing else.

These sorts of context changes can come in roughly two varieties. The first involves exogenous context changes. These are changes in the facts that constitute a moral rule that are caused by facts external to the rule. For example, take the case of a convergence on a rule that permits an elected government to take military action without a public announcement. Let us further suppose that this agreement was struck in a context where offensive military action was always a matter of ordering large troop movements that easily become publicly visible after a couple of hours. Let us suppose that Norman, a pessimist about global affairs, endorses the rule because it allows a government to more easily react to foreign enemies. Prindy, a just war advocate, also endorses the rule because sometimes a government must act in self-defence without making public announcements. But now, military action can include the bombing of far-away targets by unmanned drones without any public visibility.[16] This new context largely driven by technological advancement means that Prindy has no conclusive reason to endorse the rule anymore because the rule permits secretive offensive military action. He might think he ought to still follow the rule, but he cannot conclusively justify endorsing it. This is worrying because it is not clear whether there is any coercive power that can restrain Norman or at least decide that what he advocates is illegitimate. What society needs is an enduring social institution that can decide these issues.

The second variety of context changes involves a change in the facts that constitute a moral rule caused by the implementation of the rule itself. For example, let us say a society has converged on a rule that all public expenditure must be introduced with public savings measures of equal value. Over time let us say that this rule is followed and as a result public expenditure drops which causes the public infrastructure to depreciate in quality. This change in context will not affect all reasonable people the same way. Some reasonable people will take this as evidence that the rule is working and as such costly public expenditure has been avoided. Other reasonable people

[16] If one doubts context changes like this can have such consequences, consider the rapid escalation of the coronavirus pandemic of 2019-2020 and the way many countries without enduring social institutions were not able to form new agreements to combat the health and economic crisis. See https://nymag.com/intelligencer/2020/03/coronavirus-paid-leave-health-care-trump.html

will take this as evidence that the rule is not working and that public infrastructure requires more expenditure irrespective of public savings. In this new context, the agreement will break down purely because of its own enforcement and therefore the rule will no longer be publicly justified. Once again, this seems worrying because it would mean that there is no publicly justified rule about this issue when some rule is sorely required. What the society needs is an enduring social institution that can decide these issues.

The basic point in all of this is not that the idea of convergence itself is objectionable. Rather it is that its general application in political liberalism to achieve a stable political order by justifying issue and context specific moral rules is problematic. This is because given Diverse Packages Theory's explanation of reasonable disagreements some parts of the political order on the Convergence Conception will involve mere verbal agreements. These agreements hinge wholly on reasonable people's concepts being conveniently different to allow them to endorse the same rule. But, context changes in relation to those concepts means that agreements on issue and context specific moral rules are disposed to break down. Which means the Convergence Conception cannot show how reasonable people can achieve a stable political order because at least some of the political order it generates cannot be maintained.

3.3 Mechanism of Convergence Response

One way convergence theorists are likely to respond to the Verbal Agreement Objection is by saying that the instability identified by the Verbal Agreement Objection is a red herring. That such instability is actually a feature of the Convergence Conception. A convergent agreement breaking down in the sort of ways the Verbal Agreement Objection supposes is good because it signals that the moral rules that were publicly justified are no longer and ought not be obeyed or coercively enforced. To illustrate this point, convergence theorists will likely draw on the distinction between "stability" as ordinarily understood and "robustness", where stability as ordinarily understood is the tendency of a system to return to the same unique equilibrium point, and robustness is the tendency to return to some equilibrium point (Vallier 2019: 193–195; Gaus 2016: 230–237). The point is that the Convergence Conception shows reasonable people can maintain a political order in the robustness sense. They argue that when we model the mechanism of convergence in the right way we will see how a political order can adapt to the short-term changes in context so that it continually returns to a convergent agreement. In what follows I lay out two prominent ways this response can be made – The Multi-Perspectival Bargaining View and The Social Equilibrium

View – and argue that both have problems.

3.3.1 The Multi-Perspectival Bargaining View

Ryan Muldoon has recently argued that the mechanism of convergence should be seen as the outcome of what he calls "Multi-Perspectival Bargaining". On this view, a convergence on a set of issue and context specific moral rules is the result of reasonable people bargaining about the public rules, which specify rights, that ought to govern their shared social world. It is "multi-perspectival" in the sense that Muldoon uses the idea, developed by Scott E. Page (2007: 30–31), of people possessing and using "diverse perspectives". Muldoon (2016: 24, 64–64) argues we ought to think of reasonable people as holding different and conflicting perspectives that "provide us with filters on the world – they tell us what is important, and what we can ignore", and "shape our preferences over potential political outcomes, but they also determine what we see as the outcomes." In short, they provide the apparatus to categorise what is morally relevant in their shared social world and how they ought to judge them.

Muldoon (2016: 77–84) argues that despite holding different and conflicting perspectives, reasonable people can bargain over what rules ought to govern it because their shared social world forces them to have overlapping projections of the same states of affairs. This means although each person categorises their social world differently they overlap insofar as they are categorising the same physical state of affairs. Their "projections" are taken from this shared point. This means each recognises the stakes of having their opponent's rules governing over them. The outcome of the bargaining is a "joint individual justification" on rules over the shared social world.

Muldoon (2016: 83) argues this bargaining will take place much like bargaining in a marketplace where two parties agree to a mutually advantageous price for an exchange. In the case of issue and context specific moral rules reasonable people weigh up how much a set of rules conforms to *their* private comprehensive conceptions of justice against how much it violates it. This weighing up of the costs and benefits of the rules determines the "price" parties are willing to pay to agree to them.

Muldoon (2016: 72–77) proposes that the bargaining works like this because the object of the bargaining are rules that specify rights according to a "social conception" of rights. According to this social conception, rights come in different varieties and as "bundles of allowances and guarantees". The idea is that this is what allows people to come to an agreement about rights through bargaining because it specifies rights that have two features that allow them to be traded off against each other. One feature is that they come in different varieties like positive and negative rights, rivalrous

and non-rivalrous, and excludable and non-excludable, which means they can be extended or limited to only certain individuals in a society. Another feature is that they can be broken up into bundles of "many affiliated allowances for action or guarantees that may be considered at least somewhat independently of each other". These two features of rights, as limited and reducible to bundles of independent allowances and guarantees means that two people who do not share a perspective on how to categorise what is morally relevant can bargain their way to a set of rights over their shared social world. They are able to agree on the states of affairs they are discussing (even if these states of affairs are a small overlap of what each person sees) and on the nature of the rules that will apply to them, but not why the other person thinks the rules are right or correct.

Muldoon (2016: 90–91) acknowledges that this model of convergence does contain a measure of instability. Since the agreement is grounded in the set of diverse perspectives that reasonable people have when bargaining, the agreement breaks down either when the set of perspectives or the context changes. But, he argues, this instability can be limited to a sort that individuals can see the benefits of diversity and of striking iterative agreements which push reasonable people back into a political order without making people live by rules they do not have sufficient reason to endorse. Muldoon (2016: 102–103) argues that to achieve this, bargains about the distribution of wealth and incomes should satisfy three principles. The first principle is that the outcome of a bargain should be on the "Pareto Frontier". In short, that bargains should not only be pareto moves from some given distribution (ie. moves that can benefit a person without making someone else worse off), they should also result in pareto optimal distributions (ie. distributions where it is impossible to benefit a person more without making someone else worse off). The second principle is that the allocation of the output of a society's material production should be proportional to the contribution to that production. This means that as Muldoon (2016: 104) says, "no group should be required to subsidize another group to the extent that they would have been better off on their own." The third principle is that, with respect to the production of public goods, the state ought to permit individuals and groups to produce what those groups see as public goods as long as the burdens of production fall only on those who want them. As Muldoon (2016: 107) says, the state ought to "serve as an escrow service for groups looking to establish public goods for themselves, but with the constraint that they would have to be compatible with a bargaining process...over public goods rather than rights." The principle would then simply take the form of a requirement for another bargain made over the public goods in addition to

the one made over rights.

Muldoon (2016: 110–111) concludes that when these three principles are satisfied, this will ensure that reasonable people see the value of diversity and social experimentation. Muldoon believes that when reasonable people see that their bargains are constrained by those principles, they will see the value of living in a diverse society and of experimenting with new bargains over rights and public goods. These Millian inspired "experiments in living" as Muldoon (2016: 35) puts it, will allow people to adapt to changing contexts and learn from them. And, given they can be assured their bargains satisfy the three principles above, they will see that they are made better off in the new bargains that structure their society than backing out completely. They will see the value of experimentation and striking new bargains as a way to improve their social world.

Muldoon's account of the mechanism of convergence is compelling and does provide a plausible picture of why the Verbal Agreement Objection might not be a serious problem. If a society has, as Muldoon argues, reason to value the adaptation to changes in context and perspectives to strike new bargains, then the instability of mere verbal agreements does not seem all that worrying. But, there are I think three serious problems with the Multi-Perspectival Bargaining view of the mechanism of convergence that undermine its ability to show a political order can be created let alone maintained. The first two of which have been already made by Gerald Gaus.

The first is the Problem of Mutual Advantage. The Multi-Perspectival Bargaining way of understanding convergence makes achieving a stable political order hinge on what is mutually advantageous to the bargainers when they make an agreement. This means there is nothing in the Multi-Perspectival Bargaining view that prevents someone from backing out of an agreement on rules as soon as a rule or the "price" they paid for it is no longer advantageous to them. In fact the view encourages it as a way for people to strike new bargains. But, as Gaus (2016: 170–171) rightly points out one of the core ideas of *reasonable* people trying to agree to rules is that they seek rules that bind people despite how it might advantage them in the future. This is part of the idea of endorsing a rule as a rule that justice requires. In short, the Multi-Perspectival view seems to have an odd idea of the type of rules reasonable people are trying to propose and agree to.

The second problem is the Problem of Calculation. The Multi-Perspectival view underestimates the complexity of predicting changes in context. This complexity undermines the way reasonable people are supposed to conclusively justify the set of coercive rules by negotiating their way to an agreement. It seems almost impossible

for someone to calculate the costs and benefits of striking a bargain when they cannot predict how striking a particular bargain may change the context which will itself require a new bargain. As an example Gaus (2016: 172) refers to rules concerning immigration which affect the entry of new reasonable people into a society and therefore the make up of the set of perspectives that the constituency of reasonable people hold. But, even in a 'closed society', rules concerning education, or healthcare directly affect the context which will change the set of perspectives and therefore require a new bargain. Bargainers would need to predict all of this to have some idea of judging what rules they can agree to live under.

But aside from those two problems which have already been canvassed in the literature, there is, I submit, a much more worrying problem that has been overlooked. This is what I call the Problem of Uni-Perspectival Rights. Recall, that one of the things that is supposed to make the Multi-Perspectival Bargaining view work is that the coercive rules that people bargain over specify rights according to a "social conception" of rights.

The problem with this is precisely that it is a particular conception of rights and not a multi-perspectival one. There is no reason to think reasonable people will share such a conception because, as Diverse Packages Theory has shown that the best explanation of reasonable disagreement is that people possess and use diverse concept-conception packages. This means that when they are endorsing a rule as a rule that justice requires, there is no reason to think they will see it as specifying the kind of right that Muldoon assumes.[17] Their conception of justice could employ a completely distinct conception of rights and their concept of justice could employ a completely distinct concept of rights.

For instance, reasonable people may not share the concept of a right as merely an "allowance" and "guarantee". They may employ a more or less sophisticated idea of what a right even is. Even if there is little scope for divergence here, there will certainly be a lot of disagreement about the precise features of rights. For instance, reasonable people who believe all rights are political and institutionally defined or believe some (albeit a large portion) of rights are grounded in people's interests, will, or given by God will not agree with Muldoon's social conception of rights let alone

[17] This problem also plagues others in the public reason liberalism tradition who argue that when rights are conceived in just the right way they can in fact resolve disputes between people with diverse perspectives. See Chung (2019) on liberal rights and Chung and Kogelmann (2020: 849) on jurisdictional rights for examples where the scope of rights and one's private sphere is taken to be "reasonably" restricted, for instance by the legal system one happens to live under. It is not clear what motivates this once we account for the fact that reasonable people may not share the concept of a right.

all its features. For example, it is not clear what Muldoon's view has to say when one bargainer suggests breaking up the 'freedom of conscience' such that it does not apply to confessions to priests, but another bargainer rejects that particular freedom *can* be broken up in such a way. The point here is that the sort of bargaining that allows Muldoon's view to work relies on a conception of rights that will be the subject of the cases of deep disagreement that make up reasonable disagreement about justice itself.

All in all, the Multi-Perspectival Bargaining view is an inadequate response to the Verbal Agreement Objection. It is replete with problems running from pure practical problems of what the view requires and encourages, to a deeper problem to relying on a unique perspective on rights. This undermines the mechanism's ability to show a political order can be created let alone maintained.

3.3.2 The Social Equilibrium View

The most well developed alternative to the Multi-Perspectival Bargaining view of the mechanism of convergence is Gerald Gaus and Kevin Vallier's model of convergence as a Social Equilibrium. That is to say that convergence should be seen as an "equilibrium of social norms" (Vallier 2019: 192–195; Gaus 2016, 2011b: 434–443). The basic idea is that social norms – albeit a special set of social norms – constitute a "social-morality" that already exists in societies with reasonable disagreement about justice, and that convergence should be modelled as an equilibrium of these social norms.

Before I explain how this view works and how it might be a response to the Verbal Agreement Objection, it is worth considering what motivates it. The main motivation is that it avoids any notion of bargaining. As a result, it avoids all three of the problems faced by the Multi-Perspectival Bargaining view. An equilibrium of social norms does not ground people's reasons to endorse a coercive rule in what they find mutually advantageous given the circumstances *at the time* of agreeing, nor does it require complex calculations of the effects of an agreement. It also does not presuppose a particular conception of individual rights since the content of the publicly justified rules will depend on the social norms in a society. Given all that, the Social Equilibrium view holds out hope of modelling convergence in a way that shows it can adapt to small changes in context to achieve stability. In what follows I explain how the view works and how it faces the Conceptual Integrity Objection.

The Social Equilibrium view's model of convergence relies on two key ideas. The first is that, drawing on Cristina Bicchieri's work on social norms, we ought to see the coercive rules that reasonable people converge on as social norms.[18] The idea is

[18] See Bicchieri (2016, 2006) on the general empirical account, and Gaus (2016: 211–215, 2011b: 163–

that social norms are social rules in that they set a general standard of behaviour that is supposed to be followed by society and there is a practice of criticism and punishment for those who violate them. In a simple sense, social norms track people's reactive attitudes to certain actions in a society. Social norms with distinct moral content are social rules of a certain kind: moral rules. The distinct moral content is an ongoing mutual recognition of a structure of reciprocal deontic obligations and expectations. They are as Gaus (2011b: 181–182) says "justified deontic requirements that a large part of the population intends (at least conditionally) to follow and are actually conformed to by a large part of a group's members". That is, they are "internalized" as rules that ought to be followed by a large majority of people.

This structure has, roughly, six features which make the moral rules the sort of rules that can be publicly justified. They are sufficiently general, intelligible to all, validate claims and resolve conflicts, place requirements on behaviour not mere guidelines, are endorsable by others in different social positions, and proposed as being for the good of all such that all can reasonably internalise them (Gaus 2011b: 294–303). It is this structure that makes the social norms that are moral rules capable of being the object of public justification.

When a scheme of social norms that are moral rules are publicly justified, which means all reasonable people have sufficient moral reason to endorse them, they constitute a social-morality which establishes the "moral order" or "public moral constitution" of a society. Given this moral order is publicly justified, a political order of coercive laws and institutions is legitimate if it enforces the moral order (Vallier 2019: Ch. 6, 7; Gaus 2016: 206–207, 2011b: 449–470). This is because the political order is merely an enforcement of the moral rights that the moral rules establish.

The second key idea in the Social Equilibrium view concerns how reasonable people can identify the moral rules that all have sufficient moral reason to endorse. The idea is to see the mechanism of convergence on a set of moral rules as an equilibrium that emerges out of the actual path-dependent social interactions of reasonable people (Vallier 2019: 33–36, 110–113; Gaus 2016: 223–226, 2011b: 321–322, 389–408). Convergence – and hence public justification – is then an emergent phenomenon that arises out of reasonable people's everyday social interactions where social norms are agreed to and reassessed continuously.

This process is modelled in two stages. In the first stage, individuals rank social rules which they judge as strictly better than no authoritative rule at all. The set of social rules that all reasonable people rank in this way constitutes the *socially eligible*

182) and Vallier (2019: 30–36) on its implementation in political liberalism.

set because no rule in that set is judged as worse than no authoritative rule at all on a particular issue.[19] As Gaus (2011b: 322) says:

> The socially eligible set, then, consists in all those proposals that are unanimously ranked by all Members of the Public as strictly preferred to blameless liberty – that is, rules that all have reasons to endorse as authoritative.

To specify what it means for reasonable people to endorse a rule as authoritative, Gaus (2016: 43–44, 2011b: 38, 43, 276–279) and Vallier (2019: 4–5), like Muldoon, borrow the idea of a "perspective" from Page (2007: 30–31), but flesh it out further as "evaluative perspectives". In particular, as involving "evaluative standards" which are criteria by which rules are judged, "world features" which categorise the morally relevant aspects of one's social world, and a "mapping function" which applies the evaluative standards and world features to a proposed rule.[20] Given all that, what it means for a reasonable person to endorse a rule as authoritative involves two points.

The first point is that endorsing a rule as authoritative involves recognising, according to their evaluative perspective, that those proposing a rule have some intelligible reason for endorsing *as a moral rule* given the definition above. This means the socially eligible set will not include mere social conventions. These are conventions that people might have some reason to follow, but do not count as moral rules. Rules of etiquette would fall into this category.

The second point is that endorsing a rule as authoritative involves recognising that, according to their evaluative perspective, those proposing a rule have some reason to internalise it as a moral rule with its normative requirements and expectations. As Gaus (2011b: 322, 325) says, if a person rejects a rule they "refuse to accord it authority over him or internalize it as a rule of morality" and rules that are not rejected in this way, are "rules that all have reasons to endorse as authoritative". If reasonable people cannot even recognise the rule as such it is socially ineligible for them. As Gaus (2011b: 333) says:

> If our concern is to justify moral authority, we must suppose at some point that there is insufficient reason for according some rule (and de-

[19] I leave aside the additional step of arriving at the *optimal* socially eligible set which includes only the rules that are not pareto dominated by any other rules because convergence theorists admit for most cases the socially eligible set is enough and the additional step is irrelevant for the objection I press (Gaus 2016: 215).

[20] Gaus (2016: 51–56) adds two further features that detail an ordering of social worlds. I leave these aside for now because they are largely irrelevant for our purposes. See also Gaus (2018: 648–650) for a further development of the idea to include more features. For now the idea I summarise will suffice.

> mands based on it) authority. Such a rule manifestly fails the test of public justification. The set of rules that all Members of the Public have some reason to accept as authoritative yields what I have called a socially eligible set.

The important point here of course is not that reasonable people have sufficient moral reason to endorse any particular rule compared to the other rules in the eligible set. Rather it is that, as Gaus (2011b: 323, 325) says, the socially eligible set is a modest conclusion where if a rule is not in the socially eligible set it "fails to be publicly justified in a strong sense" such that reasonable judge those rules to "palpably fail to adequately perform their tasks." This will leave a set of moral rules about some specific issue that all reasonable people have some reason to endorse as authoritative. This means that they are willing to hold others accountable for complying with even though each reasonable person ranks the rules within that set in different ways. This is the socially eligible set.

After forming a socially eligible set, Gaus (2016: 198–202, 2011b: Ch. 4) and Vallier (2016: 202–214) argue that reasonable people must look to narrow the socially eligible set by focusing on an order of justification. This allows them to hone in on the moral rules that need to be justified first and then proceed from there. They propose that reasonable people can make use of two ideas. The first is the idea of focusing on what rights people require as moral agents. This will establish basic rights that need to be settled related to the preservation of people's status and capabilities as moral agents. The second is the idea of "jurisdictional rights" following Constant's 'liberties of the moderns'. These are rights that establish private spheres of conduct where each reasonable person's perspective is morally authoritative. These two devices provide a set of key issues that reasonable people will then be able to focus on rather than every moral rule that could be proposed.

After a socially eligible set is narrowed with those two devices, reasonable people are still left in an unsatisfactory position for creating a political order because the socially eligible set legitimates multiple incompatible moral rules. Reasonable people require some decisive way to publicly justify a unique moral rule.

To that end, the second step of public justification involves reasonable people converging on a unique moral rule – which means they find a rule they all have conclusive reason to endorse – in the socially eligible set by interacting with each other based on those rules. As such, convergence is achieved by reasonable people acting on moral rules according to 1) what moral rules *best* satisfy their evaluative perspectives, and 2) the extent to which acting on the moral rules other reasonable people are

coordinating on allows them to reap the benefits of having social interactions where they can "respect their equality and moral freedom". The benefits are that they will be able to make morally authoritative demands on others that are publicly justified. Reasonable people will then converge because there are two forces that are pushing people into agreement, the satisfaction of their evaluative perspectives and the benefit of endorsing rules that others also endorse so they can enjoy social interactions that respect other's equality and moral freedom. The point is reasonable people will be able to weigh these two considerations in such a way that they do not merely hold out for rules that *most* satisfy their evaluative perspective, but reconcile in favour of rules that satisfy their perspective as far as is possible whilst also allowing them to have social interactions with publicly justified rules. The crucial normative upshot of this whole process is, as Gaus (2011b: 414) says:

> If reasonably goodwilled people interact over a longish period of time, seeking to find mutually acceptable norms of interactions, they can come to converge on a common rule x, and this fact – that they have converged on this morality rather than that – itself can provide a public justification for x.

The idea here is that reasonable people's social interactions can be modelled in such a way to show how all reasonable people can come to have sufficient moral reason to endorse moral rules even though they started by disagreeing on what they took to be the ideal moral rules. In short, the model shows how a set of coercive rules can be in equilibrium between all reasonable people who reconcile according to their evaluative perspectives and the benefit of having social interactions with publicly justified rules.

The point in all of this is that convergence on a set of issue and context specific moral rules need not be the outcome of a procedure like Muldoon's bargaining, but an emergent outcome of people's path-dependent social interactions and revisions of social norms. As Gaus (2011b: 402–403) says:

> For once the rule is in social equilibrium (and is a recognized social norm), then all have conclusive moral reason to act on this rule rather than any other in the optimal eligible set. Thus, having created a justified rule though our interdependent choices, we can then insist that all conform to this rule, for all free and equal persons now have conclusive reason to conform to this rule, rather than any other: it is the one that best fulfills the evaluative standards of each.

It is an emergent equilibrium of social norms that can be legitimately explicated, enforced and maintained by coercive political power because they are the moral rules,

in virtue of being in equilibrium, that are publicly justified. This is how reasonable people can achieve a stable political order given reasonable disagreement.

On this view, reasonable people create a political order by freely endorsing moral rules and then constructing coercive institutions to coercively enforce them. This occurs because of the way moral rules are taken to be informal social norms that people coordinate on through social interactions they learn or adopt. An equilibrium of such social norms, when it satisfies the model Gaus and Vallier propose, will constitute an equilibrium point that describe the publicly justified moral constitution of a society. The political order is then constructed to enforce this moral constitution to either keep the unreasonable from straying or to ensure that all understand the costs of violating it.

This political order is maintained by a feedback loop between the underlying moral order of social norms that fall in and out of convergence and the coercive laws in place. This solves the Verbal Agreement Objection because it shows how changes in context will be met with changes in people's reciprocal normative obligations and expectations which will shift the equilibrium of social norms. This is because the underlying moral order is made up of social norms which form a polycentric network (Gaus 2016: 184–187). This means that rather than being a single set of principles that govern all human interactions or all social institutions, they are a network of rules that concern different behaviours at different levels of generality and in different parts of a society. This means there is no single subject, set of actions, or level of generalisation at which the social norms apply.

The polycentric nature of social norm networks allow for changes in context to be met with localised violations where people attempt to coordinate on a new rule in the socially eligible set. It allows for a practice of moral reform and criticism so that reasonable people can move to a new equilibrium of moral rules in the same process that Gaus and Vallier think the political order is created. This means reasonable people's interpersonal exchanges and moral deliberations pushe them back into an equilibrium. Given only moral rules that are in equilibrium can be coercively enforced, the political order will follow the changes in the moral order. With new coercive laws and institutions being constructed to adapt to changes in social norms.

On this view, the Verbal Agreement Objection is a red herring because the possibility of mere verbal agreements is not worrying. They are merely shifts in the equilibrium of social norms that constitutes a publicly justified moral constitution which is then enforced by a political order of coercive laws and institutions. The socially eligible set provides a set of rules that reasonable people can constantly reflect on and

deliberate about endorsing. As a result, any changes in context are part and parcel of how reasonable people converge on coercive rules.

I concede that the Social Equilibrium view does avoid the Verbal Agreement Objection. It does offer a convincing account of how the mechanism of convergence can be understood to avoid the sort of instability I argued was present in the Convergence Conception. But, I submit, all this comes at a cost, namely that the view relies on the coercive power of social norms and so faces a version of the Integrity Objection. The Integrity Objection is a well canvassed objection against the Consensus Conception of political liberalism.[21] The idea being that requiring non-liberals to only justify political principles and decisions according to the reasons they share with other reasonable people requires them to split their persona between acting as political liberals in the political domain and acting as they truly think is right according to their comprehensive conceptions of justice in the private domain. That political liberalism entails this infidelity to their true character, plans and beliefs means it attacks their integrity as reasonable people.

The Convergence Conception is, rightly, seen to avoid the Integrity objection.[22] But, I propose, there is a version of this objection that operates at the level of concepts – the Conceptual Integrity Objection – that the Social Equilibrium view of the Convergence Conception cannot avoid.[23] The idea being that modelling the mechanism of convergence as an equilibrium within the socially eligible set of social norms threatens the conceptual integrity of reasonable people. This is because achieving a stable political order depends on reasonable people endorsing a moral rule purely because of the costs of diverging from the majority who endorse it. This means that reasonable people must succumb to the social pressure of social norms to maintain a political order.

To understand how this happens, consider what Diverse Packages theory says about how reasonable people make moral and political judgements when they have deep disagreements. It says that reasonable people possess and use diverse concept-conception packages. Possessing divergent concepts of JUSTICE that input conflict considerations into their respective deliberations, and second weigh these considerations up in different ways to endorse divergent conceptions of justice. They then

[21] See Vallier (2012) for a good overview. The Objection has so far been confined to the way the Consensus Conception requires the religious to restrain using some of their core reasons to either endorse or reject coercive laws. As such it is largely about the unjustified costs restraining the sorts of conceptions of justice we ought to endorse or use (Vallier 2012: 156–160; Eberle 2002: 143–151; Wolterstorff 1997: 105).

[22] See Eberle (2011: 291–293) and Vallier (2012: 161–164) on this point.

[23] See Waldron (2015) for the closest version of the objection I press.

make judgements on, according to these conceptions, which actions are required by justice.

Now, this picture implies that the "evaluative perspectives" on the Social Equilibrium view amount to concept-conception packages. After all, they have the same function, to generate judgements about what is or is not required by justice in the world. Given that, when reasonable people have diverse evaluative perspectives what it means for a moral rule to be in the socially eligible set is that it is a rule that is supported to some extent by all reasonable people's concepts of JUSTICE.[24] This means that at least some of the considerations a person's concept of JUSTICE categorises as morally relevant support the rule. This is because the rule is seen as able to fulfil the functions of ordering a society on a particular issue in the way a rule that justice requires does. This corresponds to what convergence theorists like Gaus (2011b: 397) say about the different stages of public justification:

> The upshot of the first stage of Kantian public justification...was that *y* is eligible as a binding, moral, requirement; and according to the second, iterated interaction, stage, each Member of the Public has sufficient reason (simply given her own evaluative standards) to follow *y* over every other member of the optimal eligible set as the common binding requirement.

The Conceptual Integrity Objection concerns the way the Social Equilibrium view models how people can then endorse a single rule within that socially eligible set. Given Diverse Packages Theory, this will involve reasonable people evaluating moral rules according to what best satisfies their concept of JUSTICE *and* the benefits of coordinating on rules that others are coordinating on. This means that as more people endorse a rule the benefits of also endorsing that rule increase and therefore the balance of reasons to endorse the rule becomes weightier and eventually conclusive.

The problem with this is it entails that for some reasonable people the overriding reason that will conclusively justify a moral rule is the coercive social pressure of other reasonable people endorsing and acting on a social norm. They face others holding them accountable in their day to day social interactions. This is because at least for

[24] This point is supported by the fact that Gaus seems to accept Bicchieri's view of what it means to people to endorse social norms that are moral rules, namely that they internalise the rule by embedding it in a schema for interacting with others in varying contexts. A schema being a mental model of the appropriate normative requirements and normative expectations in a particular circumstance. But, a schema is, at least by Bicchieri and McNally's (2018: 26) description of them, a type of information structure that is like a concept, or part of a concept.

some reasonable people, the moral rule that is endorsed by others is the one that they rank as the one that least satisfies their concept of JUSTICE. Although they can see the rule as better than no rule at all, it is *barely* better. This, I submit, threatens their conceptual integrity. This is because achieving a stable political order, depends on the overriding weight of others endorsing a rule in one's balance of reasons. It does not depend on reasons related to a rule conforming accurately to what is morally relevant in their social world let alone what justice requires. Achieving a stable political order depends on at least some reasonable people foregoing acting on rules that are more integral to their concepts.

This is all very abstract. To get a sense of what the Conceptual Integrity Objection is, take a concrete example like healthcare. Consider an issue and context specific moral rule $\mathcal{M}$: Individuals and groups in the provision of healthcare have a right to discriminate against individuals on the basis of congenital illnesses and for no other reason. Suppose $\mathcal{M}$ is in the socially eligible set, meaning that all reasonable people have some reason to endorse it as authoritative according to their deliberative considerations provided by their concept of JUSTICE. Now, suppose Gordon, as part of the minority in this society, evaluates $\mathcal{M}$ as better than no rule at all because he thinks no authoritative rule at all on this issue would simply lead to even worse discrimination for individuals looking to be cared for. This would affect people finding healthcare for even common non-congenital illnesses. But, Gordon evaluates $\mathcal{M}$ as the worst possible morally authoritative rule in the socially eligible set (ie. it is closer to being ruled out by his concept of JUSTICE than any other rule). On the deliberative considerations that Gordon's concept of JUSTICE provides, he has much weightier reason to endorse all the other moral rules in the socially eligible set because they all satisfy his concept of JUSTICE better. Perhaps these are rules that each permit discrimination to a small degree, but not on a whole category of illnesses. Suppose now that a large majority in Gordon's society, contrary to Gordon, endorses $\mathcal{M}$. What is Gordon to do? What moral rule within the socially eligible set does Gordon have sufficient moral reason to endorse and coordinate on?

On the Social Equilibrium view, Gordon ought to endorse and act on $\mathcal{M}$ purely on the fact that the majority in his society endorse $\mathcal{M}$. If it weren't for that fact, he would have sufficient moral reason to endorse some other rule within the eligible set. This is not to restate the Social Equilibrium view's mechanism of convergence. Rather it is to point out that the reasonable people who are left in the minority as their society slowly converges on a single moral rule in the socially eligible set, face a stark choice. Their balance of reasons provides sufficient moral reason to endorse a rule purely on

the basis that others are endorsing it *and* that the rule is in their eligible set.

Given the way all this is supposed to be occurring through informal social interactions, it shows how creating a moral order depends on people feeling the coercive social pressure of the majority in their society acting on a moral rule they judge as barely better than no rule at all. The sole reason why Gordon will converge is that a sufficiently large number of people endorse a social norm and that joining them is more beneficial than not. Gordon ought to endorse a moral rule not because his concept represents reality accurately, but because he faces "resentment, indignation" and the "guilt" of acting on a moral rule that the majority are not converging on (Gaus 2016: 181). This threatens his conceptual integrity. He cannot, according to the Social Equilibrium view, achieve a stable political order by finding moral rules that his concepts provide sufficient moral reasons to endorse. Rather he has to also consider the fact that the majority of reasonable people in his society endorse some other moral rule according to their concepts.

One thought might be, so what? What is so bad about the threat to conceptual integrity? The main problem is that it manifestly generates its own source of instability. This is because reasonable people are likely to mistrust and hate their fellow reasonable citizens if their conceptual integrity is constantly under threat by the social pressure to endorse a social norm. This is no philosopher's flight of fancy. There is considerable evidence from political psychology that the more disagreement people face in their communicative network of political discussants, the more likely they are to become ambivalent or disengage from participating in communicative social practices within that network. For instance, Diana Mutz (2006) has argued that facing lots of disagreement generally causes people to disengage from their communicative networks where they would have the sorts of social interactions needed for convergence.[25] This can happen in two ways. Disagreement can cause a form of ambivalence where people experience a conflict within themselves about what they think and so withdraw from participating. Alternatively, disagreement when there is social accountability in the form of people having unpleasant social interactions can lead to people valuing social harmony over social interactions which lead them to change their minds or those of others. Perhaps more worryingly for the view, Nir (2011) and McClurg (2006) have backed up this evidence with studies that show when people who hold minority positions in their communicative network experience high levels of disagreement this causes them to disengage from *participating* in political procedures and not

[25] See Huckfeldt (2008; 2004) and Wojcieszak (2012; 2011) for more evidence of this, and Barnidge (2017) for further evidence of this in social media networks.

just withdrawing from "talking politics".

The worry in all of this is that there is good reason to think that reasonable people will turn their back on trying to find the moral rule that all can endorse within the socially eligible set. Any equilibrium change is disposed to fail because reasonable people are justified in either holding out for some compromise moral rule that is more supported by their concept-conception package or saying something like, "Look I cannot accept the coercive power of social pressure because necessarily it is beyond my control, but I can accept a potentially oppressive arrangement since I can at least hope to control it if I can persuade enough people". This is because they will judge the process by which social norms become publicly justified moral rules less controllable or lacking in compromise from all parties than some other method. This is a severe cost for the Social Equilibrium view. It would undermine the core mechanism by which a political order is maintained. Constant shifts in the equilibrium will appear oppressive because the path to justifying them involves succumbing to the informal costs of transgressing them as social norms.

This puts convergences theorists in a bind. On the one hand they face the Verbal Agreement Objection if they choose to be neutral on the mechanism of convergence. On the other they can accept that although the Social Equilibrium view is the best version of the Convergence Conception it comes at the cost of threatening conceptual integrity. On either side they face the problem of not being able to maintain a political order.

One response convergence theorists might make is to say that the Conceptual Integrity Objection fundamentally misunderstands what it means for a moral rule to be in the socially eligible set. That it does not mean the moral rule is supported by at least some of the considerations in all reasonable people's concept of JUSTICE. Rather it means that the moral rule is conclusively justified with respect to every reasonable person's concept of JUSTICE. This is how we should interpret what it means for a rule to be in the socially eligible set. This means it does not make sense to say that people have more or less weighty reason to endorse the rules relative to each other in the socially eligible set. This sort of view might be one way of reading convergence theorists like Gaus (2011b: 425) when they say:

> The Deliberative Model explicates the moral point of view, and what is acceptable is any option in the optimal eligible set. That is the test. If *x* is in the optimal eligible set, then *x* as a current social rule is now the basis of a moral equilibrium: a rule that has been converged upon and can be freely followed, and whose authoritative nature can be acknowledged

by each while consulting only her own evaluative standards.

On this view, a rule being in the socially eligible set does not mean that reasonable people merely have *some* reason to endorse it as I suggested earlier. Rather it means they have a conclusive reason which establishes its moral authority.

This may well be a plausible way to read the Social Equilibrium view, but it only opens it up to the worry that a stronger threshold will yield an empty socially eligible set. That, given how people can possess and use divergent concepts of JUSTICE their difference in how they interpret certain rules and rights to how they evaluate them will be deep and irreconcilable. In fact this sort of result was one of the motivations for a Convergence Conception of political liberalism. That public justification would more easily be achieved by focusing on convergence rather than consensus and by focusing on context and issue specific moral rules rather that conceptions of justice.

Another response convergence theorist might make is that, given Diverse Packages Theory's explanation of reasonable disagreement, any theory of political legitimacy that employs convergence is going to involve some threat to conceptual integrity. The depth of reasonable disagreement means that any way that a theory of political legitimacy shows how a reasonable people can conclusively justify political principles or rules is going to justify principles or rules that are some reasonable person's least justifiable option. There is no way around this. So, this is not a cost unique to the Social Equilibrium view.

I think this is far too premature. The real lesson is that the Social Equilibrium view is the best version of the Convergence Conception. But, it comes at a cost. It threatens the conceptual integrity of reasonable people and so risks undermining the core mechanism of creating and maintaining a political order in their eyes. This motivates, I submit, retaining the core idea of convergence, but exploring other types of theories. It motivates acknowledging that political liberalism is not the only way to theorise how reasonable people can have sufficient moral reason to coordinate on political principles or rules that are coercively enforced. This is what I turn to in the next chapter by exploring how conceptions of political realism fare given Diverse Packages Theory.

3.4 Conclusion

In this section I have argued that the Convergence Conception of political legitimacy cannot achieve a stable political order because it cannot show how reasonable people can maintain the order it creates. I argued it faces the Verbal Agreement Objection because an upshot of Diverse Packages Theory was that some convergent agreements

would be "mere verbal agreements". In short, I argued that convergent agreements on issue and context specific moral rules between reasonable people who diverge in their concepts of JUSTICE will be highly sensitive to changes in context. As such, at least some of the political order the Convergence Conception purports to create cannot be maintained in the face of even slight changes in social contexts.

I then argued that although there are two ways convergence theorists can model the mechanism of convergence to respond to the Verbal Agreement Objection, both of them face their own objections. I argued the Multi-Perspectival View is not an adequate version of the response because it encourages people to back out of agreements when it no longer favours them, requires complex calculations of the effect of a bargain, and requires a particular conception of individual rights that will itself be the subject of reasonable disagreement. As such, it cannot create a political order.

I then argued that the Social Equilibrium View is also not an adequate version of the response because it faces the Conceptual Integrity Objection. It relies on a mechanism that threatens the conceptual integrity of reasonable people and so generates its own sources of instability for maintaining a political order.

I concluded that although the Social Equilibrium view is the best version of the Convergence Conception it comes at the cost of threatening reasonable people's conceptual integrity. This undermines the core mechanism by which it maintains a political order. As such, the Convergence Conception is caught between the instability of the Verbal Agreement Objection, and the instability of the Conceptual Integrity Objection. Therefore, the charge that the Convergence Conception cannot achieve stability stands.

4 Conclusion

In this chapter I have argued that given Diverse Packages Theory's explanation of reasonable disagreement, political liberalism cannot provide an adequate theory of political legitimacy. This is because it does not provide a theory of political legitimacy that can show us how to achieve a stable political order. The Consensus Conception of political liberalism, as I argued in §2, cannot create a political order because it faces the Conceptual Inconclusiveness Objection and both of the ways theorists could respond to it fail. The Convergence Conception, as I argued in §3, cannot maintain a political order because it faces the Verbal Agreement Objection and both of the ways theorists could respond to it have their own problems.

The upshot of all this is that it motivates a general turn to political realism for

a theory of political legitimacy. This is the task I take up in the next chapter where I argue that extant conceptions of political realism also fail in various ways to show how reasonable people can achieve a stable political order.

Chapter 5

The Instability of Political Realism

1 Introduction

In the last chapter I argued that conceptions of political liberalism are not adequate theories of political legitimacy because they cannot show how reasonable people can achieve a stable political order. I concluded that this motivates a general turn towards political realism. The goal of this chapter is to show how, despite the fact that extant conceptions of political realism – the Non-Domination and Restrained Domination Conceptions – avoid the problems of political liberalism, they also fail to show how reasonable people can achieve a stable political order.

Recall, from Chapter 1, I said the general strategy of political realism is that it proposes that a political principle is legitimate if all reasonable people conclusively justify it as *acceptable* in a particular context. But, getting clear on the precise normative standard being used is difficult. Political realism as a first-order normative theory of political legitimacy is both heterogeneous and less developed than other approaches to political legitimacy. Theorists have argued for a political realist theory of legitimacy without clearly specifying what makes their theory a *political realist* one. With that in mind, in the same way that political liberalism has been defined by extracting the idea of public justification from its earliest proponents - largely Rawls – I will take a similar strategy with political realism.

I will take what defines political realism as a theory of political legitimacy is its use of the general normative standard Bernard Williams (2005: 3, 62–63, 135–138) calls "meeting the Basic Legitimation Demand". This is the idea that political legitimacy involves offering an "acceptable solution to the first political question". This involves justifying the political principles that are coercively enforced by an institutional structure like the modern state to secure "order, protection, safety, trust, and

the conditions of cooperation". When such political principles are justified they meet the "Basic Legitimation Demand" or BLD.

The normative standard of 'meeting the BLD' is distinguished from the political liberal's standard of 'public justification' in three ways. The first way is that the objects of justification are general political principles which prescribe the design of a society's basic institutional structure which comprises its "fundamental political framework" (Sleat 2013: 154–155; Williams 2005: 4). Unlike the political liberal standard of 'public justification' which can be applied to both political principles or context and issue specific moral rules, meeting the BLD is squarely concerned with legitimating general political principles.

The second way 'meeting the BLD' is distinguished from 'public justification' is that the nature of the conclusive justification is context-dependent. This means the facts that justify include both the content of people's moral reasons *and* the context in which reasonable people deliberate (Sleat 2013: 156; Williams 2005: 3). Again unlike public justification, meeting the BLD allows facts about the context in which people weigh reasons to justify political principles. This is understood broadly as facts about the social world, its practices, other people and their deliberations.

The third way 'meeting the BLD' is distinguished from 'public justification' is that the attitude towards the political principle that is justified is acceptance (Sleat 2013: 153; Williams 2005: 77–78, 135–138). This is a purely practical attitude towards political principles which realists claim is exemplified by the actual practice of politics. It is an attitude where reasonable people might show indignation and resentment towards the principle, but nevertheless have sufficient moral reason to comply with it. Political realists use the metaphor of games and contests that are won or lost to illustrate how one can accept the outcome of a contest, but nevertheless believe a different outcome ought to have been realised.[1] This attitude of acceptance in a contest sits between a full endorsement of a political principle as one's own and a mere compliance with it from fear of punishment. Rather, it is a matter of freely complying with a principle.

For political realists, these three features provide theories of political legitimacy that give "greater autonomy" to political thought and action (Williams 2005; Sleat 2013; Rossi 2019). It allows reasonable people to deal with their disagreements as they truly are and therefore more *realistically* than political liberalism. Different conceptions of political realism then propose different ways a political principle can meet the BLD. As we will see, political realists differ on the specific contexts that matter

[1]See Williams (2005: 13) and Sleat (2013: 139–145).

for generating conclusive justifications. But, on all conceptions of political realism a stable political order is achieved by all reasonable people's balance of reasons providing them and continuing to provide them sufficient moral reason to accept a political principle.

However, as I'll argue in this chapter, Diverse Packages Theory's explanation of reasonable disagreement about justice shows that extant conceptions of political realism cannot show how reasonable people can achieve a stable political order. To that end, the rest of this chapter proceeds as follows. In §2 I explain the Non-Domination Conception of political legitimacy, why it fails to show how reasonable people can create a political order, the responses available to political realists and my reply to each of the responses. In §3, I explain how the Restrained Domination Conception works, why it fails to show how reasonable people can create and maintain a political order, the responses available to political realists, and my reply to each of the responses. The upshot of all this is that it motivates the novel conception of political realism I propose in the next chapter.

2 Non-Domination Conception

The most prominent version of political realism is, what I call, the Non-Domination Conception. On this conception, political realism's normative standard of 'meeting the BLD' is cashed out in terms of a convergence of reasonable people's conclusive reasons produced by the normalisation of their deliberations by their shared social and cultural context. This convergence forms the grounds of stability on the Non-Domination Conception. It claims to show how all reasonable people can have sufficient moral reason to accept and continue to accept over time. In short, it shows how reasonable people can achieve a stable political order. In this section I argue that Non-Domination Conception cannot actually show this.

To that end, the rest of this section proceeds as follows. In §2.1 I lay out the Non-Domination Conception of political legitimacy. In §2.2 I explain how the conception, in light of Diverse Packages Theory, faces the Inconclusive Historical Interpretation Objection. I consider a response political realists might make and argue the response entails a descriptive theory of political legitimacy and so misses the point of this thesis which is to find a normative theory of political legitimacy. In §2.3 I explain how the conception faces the Structural Coercion Objection. I consider a response political realists might make and argue the response fails because it would rule out processes of convergence that are produced by luck or that are overdetermined by the use of co-

ercive power. I conclude the Non-Domination Conception is inadequate because it fails to show how reasonable people can create a political order given their reasonable disagreements about justice.

2.1 The Theory of Political Legitimacy

The Non-Domination Conception of political realism is best summarised as:

Non-Domination Conception: A political principle is legitimate if 1) there is a convergence of reasons amongst all reasonable people, in virtue of their interpretations of the history of their society's social and cultural circumstances, that justifies accepting it, and 2) this convergence is not produced merely by the coercive enforcement of the principle (Critical Theory Principle).

This is the theory that Bernard Williams (2005, 2002) first explicated as political realism and what other theorists have defended as a distinct conception of political realism (Horton 2010; Rossi 2013; Hall 2015; Freyenhagen 2011).[2] It relies on two core ideas. The first is that reasonable people interpreting the history of their society's social and cultural circumstances is what allows them to all find some conclusive reason to accept a political principle (Williams 2005: 11–13, 2002: 256–258). This is because their deliberations are normalised by these interpretations of the history of a shared social context. This ensures that the diversity in reasonable people's reasons and their weighing of them is reduced to the extent that their deliberations are sensitive to what political principles they have sufficient reason to accept in this shared context rather than endorsed as the ideal view of justice. The second idea is the Critical Theory Principle which states that a convergence counts only if it is not a product of the coercive enforcement of the political principle being justified (Williams 2005: 6, 89, 2002: 225–232). This avoids the cases where reasonable people live in social and cultural contexts of oppressive domination. It ensures reasonable people's deliberations are made freely rather than being merely dominated or oppressed by it. Those two ideas are what, according to Williams and those that follow him, allows all reasonable people, despite their reasonable disagreements about justice, to have sufficient moral reason to accept a political principle that is coercively enforced and therefore achieve a stable political order.

As I have already said, Williams argues reasonable people can create a political order if reasonable people form normatively rich interpretations of the history of their

[2] Arguably Judith Shklar (1989) also defends such a conception with her sensitivity to historical context and non-domination to legitimate a 'liberalism of fear'. See Forrester (2012) for a good overview of this issue.

society's social and cultural context that normalise their deliberations in a way that allows them to reach a convergence of conclusive reasons. For Williams (2005: 11) this process involves reasonable people deliberating about whether the political principle in question "makes sense" as a valid prescription for an authoritative institutional structure given their interpretation of the determinate historical facts that have shaped and constituted their society's social and cultural circumstances. This means, according to Williams (2005: 11), in light of their historical interpretations of their context, the political principle makes sense "as a legitimation of power as authority". When a political principle makes sense in this way it means the institutional structure it prescribes is "an example of the human capacity to live under an intelligible order of authority" given their interpretations. In short, it makes sense that the institutional structure it prescribes is intelligible as an authoritative political order in their interpretations. For Williams (2005: 13) the interpreting will be targeted at all sorts of facts including, but not limited to, the "obscure mixture of beliefs (many in-compatible with one another), passions, interests" that constitute the diverse conceptions of justice reasonable people hold. This sort of interpreting knits together an explanation of a society's history and allows people to evaluate whether a political principle "makes sense" in their present social and cultural circumstances.

When a political principle "makes sense" in this way to reasonable people *and* it is not the product of the coercive enforcement of the principle itself, it means reasonable people have sufficient moral reason to accept the principle in their social and cultural circumstances. When the historical interpretations are free from the influence of the coercive power enforcing the political principle it means reasonable people can be assured the process of convergence is an authentic one which grounds a *political* relationship rather than merely a form of successful domination where coercive power justifies itself. The outcome of all of this is that the historical interpretation is the grounds on which reasonable people's balance of reasons provides them sufficient moral reason to accept the political principle they are evaluating, regardless of what political principle they endorse as their own.

Reasonable people can then maintain this political order because of the way the historical interpretations are constantly updating, and also because the sort of political principles being justified. To the first point, since the creation of a political order is grounded in the interpretation of the history of reasonable people's shared social and cultural context, it means the political order will track any changes in that context. As reasonable people judge a new social and cultural circumstance as significant enough to feature in their normatively rich historical explanations, what they converge on

will also change. A society can adapt to large cultural shifts. This means that reasonable people will not be subject to political principles that were justified in social and cultural circumstances far removed from the one they currently live in. The use of coercive power under those circumstances would rightly be judged as oppressive and illegitimate.

To the second point, even though the political order tracks reasonable people's shared social and cultural context it is not vulnerable to slight changes in context. This is because the object of this justification is a general political principle for the design of an entire institutional structure like the modern state rather than issue and context specific rules as with the Convergence Conception of political liberalism. This means that what people converge upon is comprehensive enough to apply beyond a specific set of circumstances or issues. This avoids the Verbal Agreement Objection I pressed against the Convergence Conception of political liberalism in Chapter 4. Reasonable people's balance of reasons can be maintained in the face of slight changes in social and environmental circumstances since their balance of reasons supports accepting the broad terms of a political principle that is coercively enforced.

To see an example of how the theory can achieve a stable political order consider Williams's example of liberalism. For Williams (2005: 9–10) the Non-Domination Conception legitimates some form of liberalism in most western liberal democracies, if not all societies on earth. This is because we currently live in the social and cultural circumstances of "modernity". The idea is that when reasonable people interpret the determinate historical facts that constitute "modernity" – the religious wars, the prosperity of liberal institutions – they will conclude that liberal political principles make sense as a valid prescription of an authoritative institutional structure whilst a religious theocratic political principle does not. This is because a coercive institutional structure prescribed by liberal political principles is intelligible as an authoritative political order in their interpretations. As a result they have sufficient reason to accept (but crucially not endorse or adopt) a liberal political principle if they are not liberals.

But, if we consider a different society in different historical circumstances, a liberal political principle may not be legitimate. For instance, consider a deeply religious society whose political culture has been shaped by foreign invasion, the imposition of a foreign religion and a recently established liberal government. Now in this society, according to the Non-Domination Conception, moderate natively-religious political principles rather than liberalism would make sense as a valid prescription of a coercive political authority because of the historical facts that have shaped that society's social and cultural circumstances. The reasonable people of that society will have had

their ideas of self-determination, religious and political liberty shaped by the years of foreign occupation, the imposition of what they see as a false religious and recently life under liberal institutions. For them, a moderate natively-religious state would be an example of an authoritative institutional structure and therefore they would have sufficient reason to accept the natively-religious political principles even if they themselves were of a different religion.

What political principles make sense as a valid prescription of an authoritative institutional structure will vary most when comparing society's separated in time rather than distance. Societies compared at a single instance in time will vary less because historical events will most likely be shared. They will have been party to the same wars, economic crises and intellectual movements. This is why Williams (2005: 9–10) thinks some form liberalism and its political principles are the only things that make sense for most real world contemporary societies. But of course, social and cultural circumstances keep changing and people's interpretation of them will keep changing such that, whilst liberalism may be legitimate now it may not be in the future.

2.2 Inconclusive Historical Interpretations Objection

One problem with the Non-Domination Conception is that the way it proposes reasonable people can converge on a political principle through their interpretations of their society's history will be inconclusive with respect to any particular political principle. This is because as some realists have already pointed out the interpretations of a society's history will be the subject of reasonable disagreement in as much as anything else. For instance as Sleat (2013: 121–123, 2010: 489–501) has argued, William's example of "modernity" is highly contested with marxists, anarchists, and existentialists all providing their own interpretations of it. Sleat argues that "modernity" for the marxists is characterised by economic oppression and alienation, for the anarchists the growing power of the state to control and monitor individuals, and for the existentialists the rejection of religious metaphysics and a religiously defined "telos of humanity". This means the Non-Domination Conception faces a version of the Inconclusiveness Objection that beset the Consensus Conception of political liberalism: the Inconclusive Historical Interpretations Objection.

But what should really be worrying to realists is that this inconclusiveness is unlikely to be particular to "modernity" or resolved by refining our historical interpretive practices. Rather, it is what we should expect given Diverse Packages Theory as the best explanation of reasonable disagreement about justice. Diverse Packages Theory predicts that differences in historical interpretation can happen in two ways. One

way is that whilst focusing on a determinate set of historical facts they may pick on different determinate facts as more or less relevant to their interpretation. This is because they might judge different historical events as morally relevant to what justice requires given their divergent concepts of JUSTICE. Recall, from Chapter 3, this is one of the crucial functions of a person's concept. Possessing divergent concepts causes people to use different sets of morally relevant considerations in their deliberations. For instance, two people might conflict over whether the condition of working women during World War 2 was a relevant historical event for deliberating about a political principle in their current social context. This would manifestly result in people forming different and conflicting historical interpretations of their society's current social and cultural context.

Another way reasonable people differ in their historical interpretations is that they may connect the same determinate facts into a narrative in different ways. This is because they might diverge in a concept that affects the content of one of the relevant deliberative considerations provided by the concept JUSTICE. This would cause them to differ about the content of some historical event because they possess divergent concepts related to individuals, groups and ideas in those historical events. For instance, two people might conflict over the concept WORKING POOR. This would again result in reasonable people forming different and conflicting historical interpretations of their society's current social and cultural context.

When historical interpretations differ in these two ways reasonable people will have reasons filtered out in such a way that they do not have reason to accept political principles other reasonable people have reason to accept. This is because if they possess and use diverse concepts of JUSTICE, as Diverse Packages Theory predicts, then they will have different and conflict reasons filtered out in their deliberations. This will lead them to find different political principles make sense as valid prescriptions of an authoritative institutional structure. This means the core mechanism for normalising the generation of a sufficient moral reason for all reasonable people to accept a political principle fails because it is based on a consensus in justice-related judgements about history.[3] All this means that reasonable people will not all have conclusive reason to accept the same political principle. As such the Non-Domination Conception cannot show how reasonable people can create a political order.

Political realists may respond by arguing that the fact the we have seemingly stable western liberal democracies is evidence that people will form the same historical

[3]See Rossi (2013: 566–567) for a similar point about the role of consensus in the Non-Domination Conception. Although Rossi seems to see it as a feature, rather than as a bug, of the view.

interpretations of their social and cultural context. For instance, Hall (2015: 474) argues that the ability of liberal regimes to order societies in a way that they have prosperous economies without military turmoil all whilst holding political actors to account shows that liberalism is the more realistically successful way of ordering political institutions. He thinks that "even though some people will deny that liberalism in Williamsian terms makes sense, if these complaints are to be politically convincing they must offer some reasons for thinking that viable alternatives exist that will be as good at ensuring order and the conditions of cooperation here and now". Jubb (2015: 923) best sums up this view when he says, "Unless enough people find liberal political orders at least acceptable, they could not have survived". The basic idea in all of this is that the very formation and survival of liberal democracies over the last two centuries shows that we have good reason to believe people will have the same if not very similar historical interpretations despite their reasonable disagreements.

However, this response to the Inconclusive Historical Interpretations Objection does not work because it misses the point of a *normative* theory of political legitimacy. The Non-Domination Conception is proposing a theory about what makes political principles legitimate, so that reasonable people can act according to them and achieve a stable political order. But, pointing to seemingly stable states does not establish that. Rather, it gets the order of explanation backwards. We need an explanation of what makes those states stable. What Hall and Jubb would have to provide for that is empirical evidence that shows reasonable people do not differ in their interpretations of their society's history. This would provide evidence that Diverse Packages Theory does not show that people's possession and use of divergent concepts would affect how reasonable people form their historical interpretations. Moreover, western liberal democracies do not provide this kind of evidence. In many of these seemingly stable states, Diverse Packages Theory correctly predicts that there is a great deal of reasonable disagreement about the justice of a society's past. One need only look to the debates in those countries about the historical status of their indigenous peoples, the participation in past wars, and about how those countries were founded. All those debates are replete with disagreements about the justice of historical events.

2.3 Structural Coercion Objection

Another problem with the Non-Domination Conception is that even when reasonable people happen to converge in their conclusive reasons on a political principle, the conception's second condition will nearly always be violated. Most, if not all, reasonable people live within some coercive institutional structure. That is the normal social

circumstances for reasonable people. This means it is within the context of an existing institutional structure that reasonable people will form historical interpretations of their society's social and cultural context and evaluate political principles. But, engaging in historical interpretation within an existing institutional structure violates the second condition of the conception itself. This is because people will be influenced by the state's structural coercion. This means the Non-Domination Conception cannot actually create a political order.

The clearest example of this structural coercion is public education or even private education regulated by the state. These are coercive measures that will influence how people go about evaluating whether a political principle makes sense as a valid prescription of a coercive institutional structure. This is because reasonable people develop the very tools to make historical interpretations – their concepts – in compulsory schooling. The nature of that schooling is largely out of their control and coercively imposed on them. But, without that schooling they would not be able to do the sorts of things the Non-Domination Conception supposes they can do to legitimate political principles. According to the second condition all this would rule out reasonable people's process of reaching convergence.

In fact the only form of convergence that would count would be one generated completely autonomously of the state's coercive apparatus. At best this might be possible although highly unlikely given people rarely have independent access to the historical facts of their own society. At worst, it is impossible because even if there is no education related political institutions, people will still have education coercively imposed by their parents and guardians.[4] This coercion would be regulated by the state given it is the state the prescribes how much authority a parent or guardians has over their children or wards. On this view it would be impossible for reasonable people to satisfy the second condition.

In response, political realists might argue that the Non-Domination Conception's second condition should not be understood as ruling out convergence that is caused by any form of coercion. For instance, Williams (2002: 222–232), seemingly anticipating the objection, offers a more precise reading of the second condition in *Truth and Truthfulness*:

> On almost any view of the matter, however, if one comes to know that the sole reason one accepts some moral claim is that somebody's power has brought it about that one accepts it, when, further, it is in their interest that one should accept it, one will have no reason to go on accepting

[4] Thank you to Derek Ball for pointing this out.

it.

On this reading the second condition only rules out convergence caused by coercion if it is also in the *interest* of the coercive power that people converge on the political principle. This narrows the range of convergence processes the second condition rules out to only those where some coercive power influences a reasonable people's deliberations as a way to further its interest to be legitimated. This would vindicate historical interpretation as the basis of convergence because even if it is affected by the state's structural coercion it is only ruled out in case the state imposes an education regime with the express intention to ensure reasonable people converge and hence justify it. But, an education regime need not be administered this way and people can be educated from childhood under a system that coercively educates them with the intention of improving their lives generally or making them capable of evaluating political principles no matter what principles they are.

Although this response goes some way to limiting the scope of the conception's second condition, as Sleat (2013: 118–120) has argued, it is implausibly optimistic about how to assess the required counterfactuals. To see this consider two types of cases, one where a state gets lucky and one where a state overdetermines people's convergence. In the first, let us say the state has two concurrent interests: an interest that people converge on the political principle that it is attempting to enforce and an interest in enabling people to develop and use their concepts effectively when reasoning about which institutional structures to converge on. Let us say this state funds a compulsory rigorous education system funded by coercive taxes which happen to always result in reasonable people converging on the political principle that has been used to set up the institutions that comprise the state and without the education system the convergence would never happen. In short, even though it aims only at helping people to improve their reasoning and their concepts it just so happens that its coercion is crucial to being accepted. At best, the Non-Domination Conception is unclear on what to do about a case like this which would make it indeterminate. At worst, it would rule the state illegitimate which would be counterintuitive given the state was only lucky that its interest to be legitimated was further.

A second case is when the state's structural coercion overdetermines a convergence process. Let us say that the state does have an interest in it being legitimated and that it coerces to further that interest. The Non-Domination Conception would rule this out. But, let us also say that without that coercion reasonable people would converge on the political principle anyway. For instance, a society may have a long history of homeschooling that would result in children growing up to converge on a particular

political principle if there was no compulsory education regime. In this case it isn't clear why the convergence is ruled out when it is overdetermined since the coercion of compulsory education makes no difference to what reasonable people converge on.

The point in all of this is that the initial motivation of the second condition – to rule out the legitimation of oppressive coercive power – is not best served by appealing to the notion of coercion. Which political principles reasonable people accept will inevitable depend on childhood education which will be coercively imposed in one form or another. As it stands, the Non-Domination Conception, given the sorts of convergences it would rule out, cannot plausibly show how reasonable people can create a political order.

2.4 Conclusion

I have argued in this section that the Non-Domination Conception of political realism cannot show how reasonable people can achieve a stable political order because it faces two serious objections: the Inconclusive Historical Interpretations Objection and the Structural Coercion Objection. With the Inconclusive Historical Interpretations Objection I argued that the Non-Domination Conception will be inconclusive in its justification of a political principle because Diverse Packages Theory shows reasonable people possessing and using diverse concepts of justice will infect their historical interpretations. With the Structural Coercion Objection I argued the Non-Domination Conception would implausibly rule almost all political principles as illegitimate. In sum, the two objections give us good reason to think that the Non-Domination Conception cannot plausibly show reasonable people how to create a political order.

3 Restrained Domination Conception

As an alternative to the Non-Domination Conception, Matt Sleat (2013) has recently proposed a novel conception of political realism which I call the 'Restrained Domination Conception'. On this conception, political realism's normative standard of 'meeting the BLD' is cashed out in terms of a convergence of reasonable people's conclusive reasons produced by the normalisation of their deliberations by their individual contexts within a democratic procedure. This sort of convergence is what forms the grounds of stability on the Restrained Domination Conception. It claims to show how all reasonable people can have sufficient moral reason to accept a political principle. In short, it claims to show how reasonable people can achieve a stable political

order. In this section I argue that the Restrained Domination Conception cannot actually show this.

To that end, the rest of this section proceeds as follows. In §3.1 I detail the Restrained conception of political legitimacy. In §3.2 I explain how, in light of Diverse Packages Theory, the conception faces the No Simple Majority Objection. I consider a response political realists might make and argue the response runs counter to the conception's motivations. In §3.3 I explain how the conception faces the Weak Restraints Objection. I consider two response political realists might make and argue the responses undermine the motivation for the conception or, worse, undermine the motivation for theorising about political legitimacy itself. I conclude the Restrained Domination Conception is inadequate because it fails to show how reasonable people can create or maintain a political order.

3.1 The Theory of Political Legitimacy

The Restrained Domination Conception of political realism can be best summarised as:

Restrained Domination Conception: A political principle is legitimate if there is a convergence of reasons amongst reasonable people, in virtue of their deliberations as members of either the majority or minority in a democratic procedure, that conclusively justifies it as a principle to accept.

This is the sort of theory, taking inspiration from Carl Schmitt and Chantal Mouffe, Matt Sleat (2013) proposes as "liberal realism". It relies on two core ideas. The first idea is that the general political framework that underpins reasonable people's deliberation should involve a democratic political constitution. This means that the society's constitution simply *is* a democratic procedure. The second idea is the distinction between "friends", "adversaries", and "enemies" which defines the precise context that different reasonable people occupy in a democratic political constitution. Specifically, when reasonable people are friends and adversaries of a political principle they share the set of "ends, values and moral commitments" the political principle realises, whereas when they are enemies they do not. These two ideas then show when reasonable people deliberate as friends, adversaries and enemies within a democratic political constitution they can all find some conclusive reason to accept a political principle. This is because when friends and adversaries are the majority they have their ends, values and moral commitments reflected in democratic outcomes, and the enemies, despite being coerced by the majority, are coerced in a restrained way. As such the

conception shows how reasonable people can achieve a stable political order. Moreover, this improves on the Non-Domination Conception by not assuming that the convergence must rely on a consensus in historical interpretations, or that it cannot be the result of any coercion at all.

According to the theory reasonable people can create a political order despite reasonable disagreement about justice because reasonable people's deliberations are normalised by taking place under a democratic political constitution. As Sleat (2013: 169) argues, the basic political framework under which reasonable people ought to deliberate is a "political and democratic constitution". This follows the "political constitutionalism" argued for by theorists like Richard Bellamy (2007) and Jeremy Waldron (1999). The basic idea is that the constitution of the basic political framework is merely the democratic procedure of elections and representative legislative assemblies. This means all aspects of the institutional structure a political principle prescribes are implemented on the basis of democratic majoritarianism. This is opposed to "legal constitutionalism" where the constitutive elements of the institutional structure are settled in a special type of law outside the purview of ordinary legislative procedures. Examples of this would include constitutional republics or monarchies where a special form of law describes the constitutive arrangements of an institutional structure and takes precedence over the democratic decision making of a legislative body. Under political constitutionalism these constitutive elements are ordinary pieces of statue law passed by majorities in a legislature.

Within this sort of constitution, Sleat argues that reasonable people will be either "friends", "adversaries" or "enemies" of a political principle. For Sleat (2013: 153–154), "friends" are those reasonable people whose specific interpretation of their ends, values and moral commitments are realised by a political principle. In the case of "adversaries", Sleat (2013: 155) argues, that political principle only realises a plausible interpretation of their ends, values and moral commitments. In the case of "enemies" Sleat (2013: 160–164) argues, that political principle does not realise any interpretation of their ends, values and moral commitments at all.

Sleat (2013: 155–157) then argues that when the friends and adversaries of a political principle are the majority, their deliberations will necessarily provide them conclusive reason to accept it. This is because they will share a set of "ends, values and moral commitments" which given the political and democratic constitution they will pursue and realise. Friends will straightforwardly have conclusive reason because they in fact endorse the political principle. Adversaries on the other hand will have conclusive reason to accept because the principle realises at least a plausible interpretation of

their ends, values and moral commitments, and they still occupy a place within the democratic procedure of their society. This means they can participate in democratic politics and shift the institutional structure to pursue their specific ends, values and moral commitments, in which case they would become "friends" and others "adversaries".

This will leave the "enemies" of the political principle in the minority with the ends, values and moral commitments of the majority imposed on them by coercive political power. Sleat argues (2013: 160–161, 172–174) they will have conclusive reason to accept the political principle because their deliberations are normalised by recognising their context within a democratic political constitution. This means recognising that the political principle imposed on them is a result of authoritative collective decision making. The decision is authoritative because as members of the collective body that makes democratic decisions, the majority support for a political principle grounds its authority. This normalisation provides them with conclusive reason to accept the political principle even though they do not share any of the ends, values and moral commitments it realises.

But, this raises the question of why this convergence will not result in political principles that simply dominate the minority? And, as a result justify the minority of "enemies" to resist their domination. After all even though they are part of the collective decision that authorises principles it seems they do not have any influence on what those principles are. As such even though they can create a political order with their fellow reasonable people what will maintain that order?

To that end, the theory shows how this political order can be maintained with the idea that a political and democratic constitution entails the majority being committed to both a "transformative constitutionalism" and a "restrained" domination of the minority. The first idea is that in a democratic political constitution, when a political principle is enforced by the majority, it is permissible for that majority to coercively transform enemies and adversaries into friends. Sleat (2013: 158–160) explains this by adopting Stephen Macedo's notion of a "transformative constitutionalism" whereby reasonable people may permissibly maintain a political order by using public schools to shape the "intellectual development and moral character of future citizens" and restricting people's opportunities in the political sphere. For example, reasonable people may permissibly mandate that the religious publicly proclaim "the practical meaninglessness of their religious convictions as a condition of being allowed to serve." It is not hard to imagine other ways a state might try to transform the beliefs of its enemies like the use of publicly funded museums and historical exhibits or state funded

media organisations. All this is to strengthen the support for the political principle being implemented and reduce the number of enemies.

The second idea that shows how reasonable people can maintain a political order is that despite aiming to transform enemies, the friends and adversaries ought to respect them as reasonable people. This involves placing substantive limits on how the majority enforces a political principle. The majority show they are "restrained masters" by placing limits to the coercive costs involved in transforming enemies. This involves enforcing a political principle with coercive costs that are consistent with that political principle and consistent with the aim of maintaining order. According to Sleat (2013: 161–164) this will mean enforcing a political principle only by placing additional coercive costs (over and above simply being made to live according to values one doesn't agree with) on enemies when they try to actively destabilise the political order. For instance, although Sleat (2013: 163) argues that what precisely those costs can permissibly be is a "heavily contextual matter which will equally rely upon a huge degree of political judgement", he does say that some measures might include "denying them equal rights, representation, toleration, liberty". If they do not actively destabilise the political order, even through the democratic procedure, they must be treated as if they were friends and adversaries of the political principle being imposed on them. This is what maintains their balance of reasons for accepting the political principle. They are not subject to coercive costs beyond what friends and adversaries are subjected to if they do not accept society's institutions. They are respected by the majority as if they were a part of it.

All in all, the Restrained Domination Conception shows how reasonable people can achieve a stable political order by cashing out meeting the BLD with the idea of reasonable people's deliberations being normalised by their political context. Specifically a political context defined by deliberating with a political and democratic constitution and by whether they are friends, adversaries or enemies of a given political principle. The idea is that when friends and adversaries are in the majority and enemies are in the minority their balance of reasons provides them all have sufficient moral reason to accept and continue to accept a political principle.

3.2 No Simple Majority Objection

One problem with the Restrained Domination Conception is with its use of the idea of a democratic political constitution. In proposing how reasonable people can create a political order, the conception assumes a society will always contain a simple majority of friends and adversaries such that they all share a set of ends, values, and moral

commitments. But, there is no reason to assume this. Democratic procedures can plausibly involve only pluralities of people who share ends, values, and moral commitments. For instance, it is entirely plausible that a society is split into groups of 40%, 30% and 30% with each group sharing a distinct set of ends, values, and moral commitments that is incompatible with the others. In such a case the Restrained Domination Conception cannot justify any political principle and therefore cannot create a political order. The society would be in a state of political paralysis unable to actually legitimate any political principle. Under the democratic and political constitution there would always be a majority that are enemies of a political principle and therefore do not have sufficient moral reason to accept it. Any attempt to implement a political principle would be voted down by a majority. The point here is not that this will always be the case, but that the Restrained Domination has to assume it will never be the case when it is entirely plausible that the constituency of reasonable people is split into three distinct groups who are "enemies" to each other. In such a case the Restrained Domination Conception cannot achieve a stable political order because there will always be a majority who do not have sufficient moral reason to accept a political principle. This is what I call the No Simple Majority Objection.

Political realists might be tempted to say that all the Restrained Domination Conception needs is an additional condition that reasonable people have to find a set of ends, values, and moral commitments that *can* garner majority support and only the principles that realise that set are candidates for being pursued coercively. But, even this is too optimistic given Diverse Packages Theory. For that shows there is no reason to think such a majority will exist amongst reasonable people. This is because a person's ends, values, and moral commitments are determined by one's concept of JUSTICE. One's concept provides the set of values that a person then weighs up as deliberative considerations for forming beliefs about what ends and moral commitments justice requires. But, Diverse Packages Theory shows reasonable people can disagree by possessing and using divergent concepts of JUSTICE. Put simply it is entirely plausible there will be a number of concepts of JUSTICE with none shared by a simple democratic majority. There may only be concepts of JUSTICE shared by pluralities of people. As such there may not be a way to search for some very minimal set of ends, values, and moral commitments that at least a simple majority share.

A more worrying point is that this sort of response undermines the entire motivation of the Restrained Domination Conception. The additional condition on finding a simple majority is a substantive constraint on candidate political principles that tries to engineer a minimal consensus. But, one of the main motivations for the Re-

strained Domination Conceptions was to move away from relying on a consensus to achieve stability. The whole point of the general motivation to move away from the Non-Domination Conception was that, given reasonable disagreement about justice, consensus will not necessarily be found at the level of conceptions, historical interpretations or, most importantly, at the level of concepts. Simply adding a condition that reasonable people ought to look for a minimal consensus could also be used by these other conceptions in which case there would be no reason to consider the Restrained Domination Conception in the first place since we have explained away the problem that led to it. To that end, I submit the way the Restrained Domination Conception proposes reasonable people can create a political order is subject to a very plausible counterexample.

3.3 Weak Restraints Objection

Even in cases where a simple majority does exist, another problem with the Restrained Domination Conception is that the restraints on how friends and adversaries of a political principle can coercively enforce it when they are a majority, are too implausibly weak. Specifically they are weak in two crucial ways. The first way is that they permit treating a minority of reasonable people as if they were unreasonable for the sake of stability. The second way is that they permit friends and adversaries to undermine the conception's own basis for stability by moving away from majoritarianism itself. Both of these ways show the restraints on friends and adversaries are weak enough that reasonable people cannot maintain their balance of reasons to accept a political principle.

To see how the restraints are too weak in the first way, consider a case where the reasonable people who are "enemies" of the majority's political principles believe in reforming the democratic majoritarian constitution. This is perfectly plausible according to Diverse Packages Theory's explanation of reasonable disagreement about justice. The "enemies" could plausibly possess a concept of JUSTICE that includes a traditional legal constitution as an end or moral commitment in opposition to a democratic political constitution. According to the Restrained Domination Conception these enemies have to be treated as if they were friends and adversaries by being part of the democratic procedure and only being dominated (over and above having to live according to political principles they disagree with) when they actively undermine the political order.

But this is a problem because when minorities actually try to be full members of their society this will count as a threat to the stability of the institutional struc-

ture. This is because one aspect of taking part in democratic procedures is to make arguments and persuade people to try and change their minds as what ends, values and moral commitments ought to be realised. If minorities try to reform the democratic procedure itself and introduce some statutory barriers like judicial review, or some form of constitutional counter balance to the majoritarianism at the heart of the democratic and political constitution it will count as destabilising the political order. In these cases the Restrained Domination Conception permits political principles that aim to stop them for the sake of stability. For instance Sleat (2013: 162–163) says on the topic of how a state might protect its stability:

> Protecting the stability of its political framework is therefore a legitimate aim of any (legitimate) political association and can be pursued via a number of different means, part of which can often include imposing additional costs, psychological and physical, on its enemies.

But more specifically on the topic of those who threaten the basis of stability on the Restrained Domination Conception through democratic means, Sleat (2013: 163) says, "How the liberal state should respond to those who are pursuing the democratic route is going to be a heavily contextual matter which will equally rely upon a huge degree of political judgement," and that "it is an open political question whether those who seek to undermine or destabilise the liberal state via democratic means should be subject to additional costs or not." So despite enemies being treated as if they were friends and adversaries of the structure, if they *behave* as if they were by actually using their concepts to persuade their fellow citizen it will be legitimate for the structure to use coercive power to stop them. In the Restrained Domination Conception this could potentially take the form of restrictions on people's right to take part in democratic procedures or worse highly coercive measures to transform enemies into adversaries. The point here is not that the Restrained Domination Conception *guarantees* this will happen, but that its measures to prevent it are too weak. There is nothing in the Restrained Domination Conception that prevents the coercion of the state spiralling out of control once it detects minorities as threats to stability. In the end the restraints on majoritarianism are too weak such that they permit coercive measures that give enemies a reason to resist the state's coercion. It permits coercive measures that puts them in the bind of either being passive participants or being branded as threats to the stability of the political order.

To see how the restraints are too weak in the second way, consider a case where the majority of the reasonable people – the friends and adversaries – decide that majoritarian decision-making is unjust and they vote, as a majority, to move to a differ-

ent system entirely. Perhaps they decide to move to a constitutional monarchy or a super-majoritarian constitution. On the Restrained Domination Conception's own reading this is both perfectly legitimate, and yet destabilising because the basis of how reasonable people create a political order is that a majority of them support a political principle within a democratic procedure. The point here is that the Restrained Domination Conception cannot show the majority how they can sustain their own political order. They can, legitimately, at any time decide to move away from the Restrained Domination Conception's own ideas about what is sufficient for political legitimacy. This is implausibly self-undermining. This is because Restrained Domination Conception's own basis for stability – the majoritarianism and its restraints – are not justified to the majority. The point in all of this is that, even when reasonable people are treated as friends, adversaries and enemies with the restraints the Restrained Domination Conception proposes, it permits actions that do not maintain their balance of reasons.

Political realists might respond to the Weak Restraints Objection in two ways. To the problem that majoritarianism is not restrained in any way from abolishing itself, realists could respond that all we need is a substantive constraint that the majority cannot change the majoritarian basis of stability.

To the problem that the restraints on majoritarianism permit treating enemies who use democratic procedures as unreasonable, political realists could respond that in fact these are just cases where the majority do actually show restraint, but these efforts are not recognised or appreciated. Sleat (2013: 172–174) argues that the majority will have to resign themselves to the fact that some adversaries and enemies will not appreciate or experience the restraint of the state. But, showing restraint in pursing a political principle is important because the restraints explain to the majority why their rule is legitimate. The fact that the minority are not addressed by this justification or accept that these are successful justifications does not detract from the fact that these are justifications nonetheless.

But I do not think either of these responses avoids the objection. With the first response, adding a constraint that the majority cannot change the majoritarian basis of stability commits the conception to a form of legal constitutionalism which the conception was trying to avoid in the first place. The whole motivation for the democratic and political constitution was that a procedural framework could allow people to have disagreements and make authoritative decisions without any consensus on ends, values or moral commitments.

With the second response, defending the restraints on majoritarianism as actually

merely restraints on the ruling majority commits the Restrained Domination Conception to the idea that some reasonable people do not have coercive political power actually justified to them. This fundamentally concedes the motivation for offering a theory of political legitimacy in the first place. This is because it would mean treating reasonable people the same as unreasonable people which would defeat the whole motivation of trying to theorise about political legitimacy given the existence of reasonable disagreement about justice. If the restraints on domination do not justify a political principle to reasonable people then there is no sense in which they have conclusive reason to accept it. They are merely oppressed.

This speaks to the general problem of the Restrained Domination Conception's strategy of trying to weave a line between oppressive domination and non-domination using the idea of *restrained domination*. The idea is supposed to sit between compliance purely out of fear of coercion and compliance for purely non-coercion related reasons. But, one of the accepted sources of instability is the oppression of the individual by the coercion of the state and the way this coercion justifies resistance. Simply accepting that sometimes oppression is permissible undermines the whole project of theorising about political legitimacy and showing how reasonable people can achieve a stable political order. To make sense of the idea of restrained domination, political realists have to provide some account of how the restraints on domination are acceptable *to those dominated*. Otherwise, there is no distinction between it and oppressive domination.

3.4 Conclusion

I have argued in this section that the Restrained Domination Conception of political realism cannot show how reasonable people can achieve a stable political order because it faces two serious objections: the No Simple Majority Objection and the Weak Restraints Objection. With the No Simple Majority Objection I argued that there is no reason to assume, as the conception requires, that there will be a simple majority of people who share a set of ends, values, and moral commitments. As such the conception cannot show how to create a political order and would not avoid political paralysis. With the Weak Restraints Objection I argued that the restraints on how the majority of reasonable people can implement political principles are weak in two problematic ways. They permit treating the minority of reasonable people as if they were unreasonable for the sake of stability, and the majority to implement political principles that move their society away from majoritarianism itself. This means that the restraints on coercive power are too weak for reasonable people to maintain

a political order.

4 Conclusion

In this chapter I have argued that although my argument against political realism motivated a general shift to political realism, we have good reason to reject extant conceptions of it. In §2 I argued the Non-Domination Conception cannot show reasonable people how to create a political order because it depends on a consensus on historical interpretation and a constraint that forbids almost any convergent agreement reasonable people might reach. In §3 I argued that the Restrained Domination Conception cannot show reasonable people how to create a political order because it assumes that there will always be a simple majority of people who share a set of ends, values and moral commitments. I also argued it cannot show how to maintain a political order because it permits, treating reasonable people as if they were unreasonable, and the majority in a society to move away from the conception's own basis for stability.

The upshot of all this is that it motivates a new conception of political realism. One that sheds any use of consensus, and provides more plausible restraints on how political principles can be enforced. This is the task I take up in the next chapter where I sketch the Dual Convergent Conception and show how it can provide the sort of theory of political legitimacy we require, namely one that can actually achieve a stable political order.

Chapter 6

Sketch of a New Political Realism

1 Introduction

Let us step back for a moment. In the last two chapters I argued against political liberal and political realist conceptions of political legitimacy on the metric of stability. I argued that these extant conceptions either cannot show how reasonable people can create a political order or they cannot show how reasonable people can sustain a political order. There are two important lessons from this negative argument.

The first is that convergence rather than consensus provides the most promising way for a theory of political legitimacy to show reasonable people can *create* a political order. It is the best way for a theory of political legitimacy, whether of the political liberal or political realist variety, to show how reasonable people can have sufficient moral reason to coordinate on a coercively enforced political principle or rule. On balance, theories of political legitimacy have struggled the most to show how reasonable people can create a political order when they have relied on an element of consensus. This is what the Conceptual Inconclusiveness Objection and Problem of Uni-Perspectival Rights in Chapter 4, and Inconclusive Historical Interpretations Objection and No Simple Majority Objection in Chapter 5 showed. When a theory of political legitimacy uses consensus or incorporates some measure of consensus into its use of convergence it cannot show how reasonable people can create a political order. Given that, I submit, convergence understood as an agreement between reasonable people for *some* conclusive reason, is the most promising way to show how reasonable people's balance of reasons can provide them sufficient moral reason to coordinate on a political principle.

The second lesson is that political realism's normative standard is the most promising way for a theory of political legitimacy that uses convergence to show how rea-

sonable people can *maintain* a political order. This is because the three distinctive features of political realism's normative standard of 'meeting the BLD' (Basic Legitimation Demand) have proven to be the best way of avoiding the objections I made against the use of convergence in political liberalism.

Recall, political realism's normative standard of 'meeting the BLD' is different from political liberalism's 'public justification' in three ways. The object of justification is a *political principle* as opposed to issue and context specific rules. The facts that conclusive justification depends on are the content of people's moral reasons and the *context* in which they deliberate. The attitude that is justified is to practically *accept* a political principle as opposed to endorsing it as a true principle of justice or moral rule.

Those three features avoid the Verbal Agreement Objection and the Conceptual Integrity Objection I made in Chapter 4. Allowing reasonable people's context to normalise their deliberation meant that reasonable people with highly diverse sets of moral reasons could still find a way to agree given their shared context. Focusing on merely justifying acceptance as opposed to endorsement meant that the threshold for any individual to reach conclusive justification was lowered. Both features made it possible for conceptions of political realism to show a reasonable could then converge on something as general as political principles. All this avoided the Verbal Agreement Objection because it showed how political principles that realise conceptions of justice comprehensive enough to deal with changing social and environmental contexts could be justified. It also avoided the Conceptual Integrity Objection because it shows how reasonable could converge and accept a coercively enforced political principle without having to give up using the concepts they believe accurately represent their social world. Given that, I submit, political realism's normative standard is the most promising way for a theory of political legitimacy that uses convergence to show reasonable people can *maintain* a political order.

These lessons show that what's needed is a political realist theory of political legitimacy that commits more thoroughly to convergence. A theory which, unlike the Non-Domination and Restrained Domination Conceptions, sheds any use of consensus, but makes use of the political realist normative standard. With that in mind, in this chapter, I propose a novel conception of political realism – the Dual Convergent Conception. The theory combines elements of the political liberal's social equilibrium version of convergence with political realism's normative standard. The next section details this theory and the following two sections consider and respond to two objections that could be made against it.

2 Dual Convergent Conception

The general strategy of the Dual Convergent Conception is to combine the political realist normative standard of 'meeting the BLD' with the insights from political liberals like Gerald Gaus on the mechanism of convergent agreements. The way I propose to do this is with the central organising idea of 'ordered moral warfare'. This is a state of affairs constituted by two convergent agreements, or a 'dual convergence'. First, a convergence on a political norm that prescribes a procedure for selecting a political principle to be coercively enforced, and second on a set of political principles that concern the design of a society's basic structure of social institutions. In the following subsections, I detail what the idea of ordered moral warfare is, how it achieves a stable political order, and its comparative advantages over other conceptions of political realism.

2.1 The Idea of Ordered Moral Warfare

By a "political norm that prescribes a procedure for selecting a political principle" I have in mind a particular type of social norm, which prescribes a particular type of activity. The particular type of social norm I have in mind are the norms that Jon Elster (2014: 53) calls "strategic norms" which constitute a coordination equilibrium where all conform to the norm as the best rational response to how others act or diversity in general. However, unlike Elster the political norm I have in mind is not left to "the closed circles of government and parliament" or without any moral content. Rather they share a feature of what Elster calls "non-strategic norms" in that they are unwritten rules that "regulate the basic machinery of politics" which are enforced by the citizenry at large for moral reasons.[1] The best way to think about the type of political norm I have in mind is as the sort of social norm described by political liberals like Gaus (2016: 180–183) which involves a practice of accountability where there is an ongoing recognition of a structure of reciprocal normative obligations and expectations in relation to some behaviour.

However, the political norm that is part of the idea of ordered moral warfare is not a norm about any social or political activity. Rather it is more limited in scope. It is about how a society ought to transition from some coercively enforced political principle to another. In short, they are unwritten rules that prescribe a procedure for socially selecting *some* political principle (I will say more about what these principles are

[1] It should be plain here that the political norms I have in mind are the sort of mixed cases that Elster (2014: 59) mentions in passing.

soon). For most societies this political norm will prescribe some collective decision-making procedure like representative democratic voting, or hierarchical council decisions, or even nation-wide referendums. But, the key in all these procedures is that their ultimate normative force and coercive enforcement does not depend on their place in a written constitution or enforcement by the current political power. That would merely lead to a regression in explaining how the particular procedure is justified because the question would always arise about how the written constitution was justified. Given that, the procedure's justification depends on a political norm's unwritten structure of obligations and expectations demanding that people participate in that procedure, select a political principle with it, and comply with its results. Although such a norm may then be codified as law to reinforce the practice of accountability, this fact is not what its normative force or coercive enforcement ultimately depends on. Rather it would depend on the political norm's unwritten structure of obligations and expectations conferring moral significance on *that* piece of law as a law that ought to be obeyed. That is why the collective decision-making procedures are ultimately enforced by citizens at large. They are enforced by citizens rebuking each other for not participating, for not complying with its results or in the most extreme circumstance violently enforcing the norm when they are violated. When the political norm is codified as law, its enforcement by the political authority is contingent on the fact that there is a convergence on the political norm which prescribes the procedure that selected the political principle that the political authority acts on.

By a "set of political principles" I mean a set of general principles for the design of a society's basic structure of social institutions. These principles may be a part of conceptions of justice, or they might be discreet principles related to specific institutional actions. The important point is that they are not issue and context specific moral rules like the social norms that featured in the Social Equilibrium view of convergence from Chapter 4. They are general principles that for many people may be part of a conception of justice that designs society's basic social institutions.

Now, the heart of the idea of ordered moral warfare is that a convergent agreement on a particular political norm that prescribes a procedure for selecting a political principle *and* a convergent agreement on a particular set of political principles, together allow for a unique political principle that is a member of that set to be conclusively justified to all reasonable people. This is because the convergent agreements involve for all reasonable people the conclusive justification of a political norm and the conclusive justification of a set of principles, respectively. Individually, the justifications of the set of principles and the political norm only provide *pro tanto* justifica-

tion of a particular political principle. But, together the two convergent agreements provide two reasons that together *conclusively justify* accepting a particular political principle. The agreements are then jointly sufficient for reasonable people's balance of reasons providing them and continuing to provide them over time, sufficient moral reason to accept a political principle. This is because the particular political principle is conclusively justified based on the conclusive justification of the set of principles of which it is a member and the political norm that prescribes the procedure that selected it out of that set.

In effect this is a general form of proceduralism where a particular political principle is justified based on the fact that it is the output of some procedure for selecting political principles. Except in our case, what justifies that output is not that the procedure embodies some morally relevant property of fairness, reasonableness, or rationality. But rather, that the output is the result of a procedure prescribed by a political norm that all reasonable people have converged on, *and* that the output is within a set of political principles that all have converged upon. The fact of those two convergent agreements are what justify a particular political principle when it is selected.

But of course that simply raises the question of how conclusive justification is achieved in those two convergent agreements. I propose that the facts that conclusive justification depends on in the two agreements are cashed out in terms of a restriction on what counts as a relevant moral reason and a context that normalises the deliberation of those reasons. The restriction on what counts as a relevant moral reason, like the Social Equilibrium view of convergence, is a weak restriction on the content of the reasons that reasonable people's concept of JUSTICE provides for deliberating about political norms and political principles. The restriction is that people's reasons ought to be mutually intelligible to others as a *moral reason* for the person who possesses it. This involves a mutual recognition that a person's moral reasons makes sense as a moral reason given that person's concept of JUSTICE. This excludes reasons that reasonable people might profess to have that are completely out of character or unrelated to justice.

The context that normalises the deliberation of the mutually intelligible moral reasons varies depending on the convergent agreement people are trying to reach. When reasonable people converge on a political norm, the context that normalises their deliberation is the historical status quo political principle they live under, and the political principles that are actually advocated for in their particular society. In short, the context is the political culture of their society. However, when reasonable people converge on a set of political principles, the context that normalises their de-

liberation is their earlier convergent agreement on a political norm. Although all this will be fleshed out in more detail later, for now it suffices to illustrate the way contexts will be one of the facts that conclusive justification depends on.

The result of conclusive justification depending on those facts is that it models how reasonable people's balance of reasons provides, what I call, 'compliance-for-the-right-reasons'. This is compliance not out of fear of coercion, the restraints on domination, or what makes sense given some interpretation of the history of a society's social and cultural circumstances. Rather it is compliance from the moral reasons their concept of JUSTICE provides within the context of a society's political culture, and the context of the convergent agreement upon a particular political norm. All this yields the following theory:

Dual Convergent Conception: A political principle is legitimate if 1) it is selected by a procedure prescribed by a political norm that a convergence of mutually intelligible reasons amongst all reasonable people and the context of their political culture, conclusively justifies accepting, and 2) it is a member of a set of political principles that a convergence of mutually intelligible reasons amongst reasonable people and their context of having converged on a particular political norm, conclusively justifies accepting.

The two convergent agreements together constitute a state of affairs of 'ordered moral warfare'. This is a state of affairs, contra political liberals like Gaus (2016) and Vallier (2019), where politics is a form of warfare where reasonable people constantly attempt to change which political principles are selected by acting according to a political norm for that very purpose. As such, reasonable people do not live under the utopian goal of "moral peace" where, according to Vallier (2019: 2–3), people live in a society "with a high degree of justified social trust". Rather, they live in a form of warfare that is ordered within the bounds of morality because reasonable people's balance of reasons provides compliance-for-the-right-reasons. Neither does the theory aim to describe society's "public moral constitution" which according to Gaus (2016: 177–180) is the basic framework of moral rules that underlies reasonable people's shared social world. Rather, the theory explains the political legitimacy of a coercively enforced political principle such that reasonable people can achieve a stable political order by acting on it. To that end, the two conditions jointly show how reasonable people's balance of reasons can provide them, and continue to provide them over time, sufficient moral reason to accept such a political principle. In short, the convergent agreements are jointly sufficient for the legitimacy of a political principle.

2.2 Achieving a Stable Political Order

As I have said, what shows reasonable people can achieve a stable political order on the Dual Convergent Conception is the idea of ordered moral warfare. This idea is constituted by two convergent agreements and a particular sort of deliberation that models how reasonable people can reach those agreements. The idea shows how reasonable people can create a political order despite reasonable disagreement about justice by cashing out reasonable people's conclusive justification to accept a political principle as a combination of two separate justifications. These two separate justifications are the conclusive justification to accept a political norm that prescribes a procedure for selecting a political principle and the conclusive justification to accept a set of political principles.

These conclusive justifications are modelled in a similar way to how political liberals like Gaus (2016: 208–226) model convergent agreements on social norms that are moral rules. We start with a deliberative model in which reasonable people's conclusive justifications are modelled by their deliberations about coordinating on political norms and political principles. Beyond this, the Dual Convergent Conception diverges in three important ways. The first way is that the political norms and political principles they deliberate about are those that are openly endorsed presently and in the history of their society's political culture. This does not mean that they need to know all the norms and principles every held by any reasonable person. Rather, they need only consider those that have been openly advocated for and therefore brought to their attention by their fellow citizens. The second way is that reasonable people's deliberations are affected by both the reasons provided by their concepts of JUSTICE *and* the social contexts they occupy when deliberating. The third way is that the attitude that is conclusively justified is *acceptance* and not *endorsement*. Although there is no lexical priority to the agreements, there is a causal priority and so I will begin with explaining the convergent agreement on the political norm first.

The convergent agreement on a political norm that prescribes a procedure for selecting political principles is modelled, as in Gaus's model, in terms of comparative preferences. This means it is modelled in terms of the political norms reasonable people prefer over a justificatory baseline (Gaus 2011b: 304–310). However, since I aim to model how people have conclusive reason to *accept* a political norm rather than *endorse* it as their own, there will be two justificatory baselines. The preferences over these two justificatory baselines are cashed out in terms of two choices that reasonable people make about political norms. The idea is that these two choices will model the way the reasons a reasonable person's concept of JUSTICE provides are normalised by

a shared social context. This normalisation is what allows reasonable people with diverse concepts of JUSTICE to make the deliberations necessary to converge and accept, rather than endorse, a political norm.

The first choice people make is to ask themselves: "Given the historical status quo political arrangement, which political norms do I prefer over a political norm that enforces a procedure of remaining at the status quo for the forseeable future?". This yields a set of political norms that reasonable people prefer over the uncertainties that come with living in an unjustified political order. This is a preference not over a chaotic state of nature or state of "blameless liberty" as Gaus (Gaus 2011b: 322) has in his model. Rather it is a preference over an unjustified status quo order that may be overthrown or continue by oppression without justification. As such, it is a preference over the uncertainties of living under a political order that all do not have conclusive reason to coordinate on.

The second choice people make is to ask themselves: "Given the set of political principles openly endorsed now and in the history of my society's political culture, which political norms do I prefer over a political norm that enforces a procedure of randomly selecting political principles?" This further narrows the set of political norms from the first choice, to a set of political norms that reasonable people prefer over the uncertainty of having to live under political principles merely selected at random. It models a preference over having no control of the political order one lives under even though it would be justified.

The outcome of both choices is, I submit, a set of political norms that a reasonable person has conclusive reason to accept. Importantly, this does not mean they have conclusive reason to accept a particular political norm over all others in the set. Rather, it means they have conclusive reason to accept any of the political norms in the set over all the other political norms openly endorsed presently and in the history of their society's political culture. This is because they have evaluated these norms according to their concept of JUSTICE *and* the social contexts mentioned in the antecedent of the two choices above. As such, the set only includes those norms that a reasonable person could coordinate on over political norms that enforce the unjustified status quo political order, or enforce a procedure gives them no control over whether they live in a justified or unjustified political order. In short, the choices yield the set of political norms that a reasonable person is willing to reconcile on given they value living in a stable political order and value achieving it by controlling their social world rather than by pure luck.

To move beyond the individual perspective and model how *all* reasonable people

can converge on a political norm we then take the overlap of every reasonable person's individual set. This overlap yields the set of political norms that *all* reasonable people have conclusive reason to accept, or as I will call it the *social acceptance set of norms*. Of course, every reasonable person will have a personal ranking of the political norms within the *social acceptance set of norms*. This is because they are not indifferent *between* the political norms in the *social acceptance set of norms*. But, even if the strength of the reason to accept varies between the norms, each individual person has at least pro tanto reason to accept every single norm in the set over any other political norms. This is what it means for the set to be the set of political norms that *all* reasonable people have conclusive reason to accept.

But of course, a set of political norms that all have conclusive reason to accept is not enough. To avoid disorder what reasonable people require is to coordinate on a single norm and therefore conclusive reason to accept a particular political norm in the *social acceptance set of norms* over all others. Here again political realism can take a cue from political liberals like Gaus. The model supposes that reasonable people will recognise that some choice needs to be made within the *social acceptance set of norms* and so they will converge through their path-dependent social interactions that are particular to their society and its history. There are many ways this can be specifically modelled, ranging from 2-person iterative impure coordination games, to N-person iterative impure coordination games within a single generation.[2] However the convergence is modelled, in reality it will be an emergent social evolutionary phenomenon. It will involve many historically extended social interactions where convergence on a single political norm will depend on socially contingent events in a society's history that relate to political norms. For example, revolutions, civil wars, constitutional conventions, high court decisions, referendums and even the publishing of philosophical works will all affect how reasonable people in a particular society will weigh the benefits of coordinating with the political norm others are coordinating on. At the end of all this every reasonable people can have some conclusively reason for accepting the same political norm despite their reasonable disagreements about the ideal political principles.

With that explanation of the convergent agreement on a political norm in place, we are in a position to see how the convergent agreement on a set of political principles is supposed to work. This convergence is modelled in much the same way as the convergent agreement on a political norm. It is modelled in terms of the political

[2] See Gaus (2011b: Ch. 7), and Vanderschraaf and Skyrms (2003) on these models. See also Shotter and Sopher (2003) on empirical evidence for these results.

principles people prefer over two justificatory baselines. Except, now there are two important differences. The first difference is that the object of justification is a set of political principles out of the political principles openly endorsed presently and in the history of their society's political culture. The second difference is that people's deliberations are modelled in terms of two choices reasonable people make in light of the social context of the convergent equilibrium on a particular political norm they have already reached. In short, the convergent agreement on a political norm about the procedure for selecting a political principle becomes a shared context that normalises the reasons reasonable people's concept of JUSTICE provides. This normalisation is what allows them to make the deliberations necessary to converge on a set of political principles.

The first choice they make is to ask themselves: "If the political norm we have converged on enforces a procedure that *does not* select the political principle I conclusively endorse, then which political principles do I prefer over any possible random principle?". The answer to that question yields the set of principles an individual prefers over the uncertainty as to which political principles will be selected, given the certainty it will not be the one a person conclusively endorses. This is not a preference over a state of nature, or state of no justified political principles at all. Rather, it is a preference over simply having some principle being selected with no regard for whether it conforms to one's concept of JUSTICE or not. It is a preference over having no control of how close one's society can approach the political principles one conclusively endorses.

The second choice reasonable people make is to ask themselves: "If the political norm we have converged on enforces a procedure that *does* select the political principle I conclusively endorse, then which political principles do I prefer over any possible random principle?". The answer to that question further narrows the set of principles to the principles an individual prefers over the uncertainty as to which political principles *could possibly* be selected, given the principle they conclusively endorse is currently selected. This, again, is not a preference over a society returning to a state of nature, or state of no justified political principles at all. It is a preference over having no control of how far one's society could stray from the political principles one conclusively endorses.

The outcome of both choices is, for every reasonable person a set of political principles they have conclusive reason to accept. Importantly, this does not mean they have conclusive reason to accept any particular principle in the set over all others. Rather, it means they have conclusive reason to accept any of the political principles

in the set over all the other principles openly endorsed presently and in the history of their society's political culture. This is because they have evaluated these principles according to their concept of JUSTICE *and* the social contexts mentioned in the antecedent of the two choices above. As such, the set will only include principles that, a reasonable person could coordinate on whether her favoured ideal political principle – the one she has conclusive reason to *endorse* – is or is not selected. In short, the choices yield the set of principles that a reasonable person is willing to reconcile on.

However, much like the convergence on political norms, to move beyond the individual perspective and model how all reasonable people can converge on a particular political principle we first take the overlap of every reasonable person's individual set. This yields the *social acceptance set of principles*, which is the set of principles that *all* reasonable people have conclusive reason to accept. But of course, every reasonable person will have a personal ranking of the principles within in the *social acceptance set of principles*. They are not indifferent between the political principles given they are evaluating them based on their concepts of JUSTICE *and* the contexts mentioned in the two choices above. Nevertheless, even if the strength of the reason to accept varies between the principles, each individual person has at least pro tanto reason to accept every single principle in the *social acceptance set of principles*. This is what it means for the set to be the set of political principle that *all* reasonable people have conclusive reason to accept.

But of course, a *set* of political principles that all have conclusive reason to accept is not enough. To actually create a stable political order reasonable people need to coordinate on a particular political principle. Therefore they need to have conclusive reason to accept a particular political principle in the *social acceptance set of principles* over all others. Here, on the Dual Convergent Conception, political realists need not rely on any social evolutionary mechanism. Rather, we can rely on the earlier convergence on a political norm. It is the fact that a particular political principle within the *social acceptance set of principles* is selected by the procedure prescribed by the converged upon political norm, that then *conclusively justifies* that particular political principle.

The obvious question at this stage is, why should we think the *social acceptance set of principles* will not be a null set? Why will reasonable people not hold out and judge that in both of the choices they do not prefer any principles other than those they conclusively endorse? Why should we think reasonable people will be pushed to reconcile? I propose two reasons.

The first reason is that reasonable people's deliberations will be normalised by

their shared social context. Reasonable people will be pushed to converge rather than diverge in the deliberative process because they are not relying purely on the considerations their concepts of JUSTICE categorise as morally relevant. Rather they also rely on the shared social context of having converged on a particular political norm that prescribes a procedure for selecting political principles. This social context helps reasonable people converge by providing information relevant for constraining their deliberation.

One sort of information the social context provides is about which principles are consistent with and support the ideals of citizenship that reinforce the political norm they have converged on. This means that reasonable people will have some shared reasons to accept or reject principles. But importantly they will not share them because they have the same concept of JUSTICE. They will share the reasons because they can recognise how certain principles will conflict with or undermine the political norm they have converged on. Such principles would undermine an equilibrium they as reasonable people have arrived at.

Another sort of information the social context provides is about which political principles are more or less likely to be selected. This means that reasonable people can construct shared predictions of what kind of social world they would end up with when making their choices. Combined with their own concept of JUSTICE, reasonable people can evaluate the value of accepting political principles that range across the spectrum of easy to realise, to too difficult to realise. For instance, if the procedure the political norm prescribes is "yearly referenda", then a person can predict which political principles are more or less likely to be selected. They can make predictions according to the way the questions on the referenda ballot are decided and given the political principle their fellow citizens advocate for. These judgements will constitute a shared set of predictive models of which political principles are likely to be selected and which are not. These models will help reasonable people weigh their considerations to a sufficiently similar degree. It encourages them to reconcile towards political principles that are more likely to be selected than not, but also towards those that do not conflict with their concept of JUSTICE so much that they do not have reason to accept it.

Aside from the shared information, the second reason reasonable people will be pushed to reconcile is that as reasonable people they can recognise two facts about their social and biological reality that underpin their political life. These social facts have largely been ignored in contemporary political philosophy because of the propensity to idealise reasonable people too much away from actual people. How-

ever, this ignores the 'production' and 'decline' of reasonable people. These facts are important if a theory of political legitimacy is about how actual reasonable people ought to act.

The first social fact is that reasonable people do not emerge into existence fully formed as reasonable people. They are initially children who live a largely non-autonomous life. To develop into reasonable people they depend on other autonomous people to develop their capacity for a conception of justice. Experimental evidence shows that children innately possess moral concepts, but these remain crude and applied in highly localised ways. More global judgements that transcend "the enemy of my enemy is my friend" evaluations require socialisation to develop these innate moral concepts into the sort that can be of use in political life.[3] Reasonable people cannot develop their conception of justice without depending on others to develop their innate conceptual tools. Children require socialisation and care to develop their moral concepts of the right and the good to then go on to become reasonable people. Reasonable people are forced to recognise the social interactions that satisfy this sort of dependence is itself a good. This is because neither they nor anyone could have a conception of justice without it.

The second social fact is that reasonable people as adults do not maintain their capacity for a conception of justice forever. Even though they might hold on to their capacity for a conception of the right, they require, much like children, socialisation to maintain their capacity for adopting and exercising ideas of what is good for them and others. Experimental evidence shows lower political participation, general cognitive decline, declining physical and mental health, and most importantly not being able to conceive of what is good in one's social world, are all linked to social isolation.[4] Adults require others to take an interest in how their lives turn out and their views of the good, to maintain their capacity for a conception of the good. As such, reasonable people are forced to recognise the social interactions that satisfy this sort of dependence is itself a good. This is because they could not sufficiently maintain their capacity for a conception of the good without it.

Recognising the two social facts I have sketched forces reasonable people to recognise a unique good of political legitimacy: the good of a social union of mutual dependence. This is a social union in which reasonable people are able to depend on those they disagree with for developing and maintaining their capacity for a conception of

[3] See Railton (2017), Hamlin (2015: 504–506, 2017), and Nucci et al. (2017). See also Bloom and Wynn (2016) for a good overview of the issue and further evidence.

[4] See Reilly (2017) and Evans et al. (2018), and most importantly Cacioppo and Patrick (2009: 14–15, 100–108, 180–181).

justice. As such, it embodies a deep form of reciprocity where each reasonable person makes the development and maintenance of each people's capacity for a conception of justice part of their conception of the good. This is a good that all reasonable people can recognise. This is because it is a social union that allows them and those they care about to have the social interactions that help them develop and maintain their capacity for a conception of justice.

Recognising the good of a social union of mutual dependence gives reasonable people reason to reconcile. It is only by reconciling that reasonable people avoid a scenario where they and those they care about fail to develop and maintain their capacity for a conception of justice. To depend on others to help those we care about to develop and maintain a capacity for a conception of justice requires that one is willing to coordinate on political principles that do not realise one's own conception of justice. This is because we cannot expect those who disagree with us to make the development and maintenance of people's capacities part of their conception of the good if we do not at least value their ability to realise their conception of justice. We must be ready to see how we could coordinate on political principles we do not endorse. This is because coordinating on political principles that realise other people's conceptions is the only rational way to expect them to help us when we require it.

Likewise, for others to coordinate on the political principles that realise our conception of justice requires us to develop and maintain the capacity for conceptions of justice in the people that other reasonable people care about. This is the case even though we can be assured these other people may settle on conceptions we disagree with. This means we must be ready to reconcile by tolerating the development of capacities in others who we disagree with. This is the only way to expect others to reconcile and coordinate on the political principles that realise our conception of justice.

In sum, reasonable people will reconcile because they recognise the good of a social union of mutual dependence. This reconciliation will then, in the model I have sketched, provide them with sufficient moral reason to accept a political principle and therefore create a political order. This is because, despite the depth and breath of reasonable disagreement about justice, each reasonable person will have a mixture of two reasons that together provides them conclusive reason to accept a political principle. They will have a conclusive reason to accept the political norm that prescribes the procedure that selects a political principle, and a conclusive reason to accept the set of principles from which the procedure selects a principle. Taken together, when the procedure prescribed by the political norm, which reasonable people have converged on, selects a political principle that is a member of the set of principles, which they

have also converged on, all reasonable people have sufficient moral reason to accept that political principle. To that end, the theory shows how reasonable people can create a political order.

But as we have seen throughout this thesis creating a political order is not enough. A theory of political legitimacy also has to show reasonable people how to maintain a political order. The Dual Convergent Conception shows this through three features inherent to the idea of ordered moral warfare. These features show how the political order the theory creates can resist the endogenous and exogenous forces that disturb reasonable people's balance of reasons. They show how reasonable people can continue to have sufficient moral reason to accept the political principle over time.

The first feature is that the object of justification is a general political principle rather than merely a context and issue specific moral rule. This means the convergence on it is resistant to slight changes in social and environmental circumstances. Convergence on a political principle involves a convergence on a general view of how to construct social institutions. This sort of convergence will involve considerations on various issues and across various contexts. This avoids the Verbal Agreement Objection that plagued the use of convergence in political liberalism. This is because what is converged on is meant to apply across various contexts. Slight changes in context do not cause the agreement to break down.

The second feature is that the convergent agreement on the political norm is reinforced every time a political principle is selected by the procedure prescribed by the political norm. This is because the institutional structure the principle realises will seek to promote and strengthen reasonable people's agreement on the political norm. Those who conclusively endorse the political principle will seek to endorse certain ideals of citizenship that will promote compliance with the political norm. This of course does not mean they will enforce a political norm no matter what occurs or how much the social and environment context related to the political norm changes. Rather it merely adds a degree of rigidity to the convergent agreement on the political norm.

The third feature, is that the convergent agreement on the set of political principles involves an inherent toleration of other people's concepts of JUSTICE. This means none of the principles, once realised, produce their own instability. This is because one of the choices that modelled people's deliberations was which political principles they prefer over any random principle when the principle they conclusively endorse is selected by the procedure prescribed by the political norm. This means that one of the considerations that feature in people's deliberations is the principles they could tolerate when the principle they endorse according to their concept of JUSTICE

is deselected and their society selects another competing principle. In simple terms they had to deliberate about which principles they could tolerate once they had 'won the political contest' the political norm constitutes. When all reasonable people deliberate in this way, the set of political principles they all have conclusive reason to accept is a set that all can tolerate and therefore will not aim to rule out or erase when their political principle is selected. This allows reasonable people to exercise their autonomy to use the concepts they believe represent the world accurately. They can categorise their social world according to their concepts and advocate for the political principles they justify endorsing. When a political principle they do not endorse is selected by the procedure prescribed by the political norm, they can be assured they do not have to endorse it as their own, or that their advocacy for the political principles they do endorse will not be suppressed. They merely have to practically accept the political principle whilst still being permitted to argue against its implementation and advocate for their preferred political principles.

In sum, the Dual Convergent Conception shows how reasonable people can achieve a stable political order with the idea of ordered moral warfare. That central organising idea is constituted by two convergent agreements. The way deliberation is modelled in each and the particular way the agreements are related, show how reasonable people can have and continue to have sufficient moral reason to accept a political principle that is coercively enforced.

2.3 Comparative Advantages

Importantly the way the Dual Convergent Conception achieves a stable political order has a number of comparative advantages over other conceptions of political realism. On the first advantage, recall the Inconclusive Historical Interpretation Objection and the No Simple Majority Objection. The former was that the Non-Domination Conception relied on a consensus in reasonable people's historical interpretations. The latter was that the Restrained Domination Conception relied on a consensus in reasonable people's concepts of JUSTICE to the extent there was always at least a majority of people who shared a set of ends, values and moral commitments. But, there was no reason to think that either form of consensus would exist between reasonable people. As such I concluded that both the Non-Domination Conception and Restrained Domination Conception could not show how reasonable could create a political order. However, the Dual Convergent Conception avoids both of those objections. It does not employ any degree of consensus on a historical interpretation or on justice. Rather it only requires reasonable people recognise a type of social

union that is necessary for them and those they care about to develop and maintain a conception of justice, as a good.

On the second advantage, recall the Structural Coercion Objection against the Non-Domination Conception. The crux of the objection was the idea that if one forbids any convergence that is caused by coercion, as the Non-Domination Conception does, this would rule almost all political principles or rules illegitimate. This would then mean that no political order, or almost no political order, could ever be created. However, the Dual Convergent Conception does not apply any constraints on how convergence may permissibly be produced that is linked to coercion. Rather it only requires that reasonable people have conclusive reason to accept a political norm and a set of political principles on the basis of their concepts and their social context. This accepts that reasonable people's social worlds are complex systems where avoiding any form of coercion is impossible, but nevertheless grounds the justification of political principles in the content of their moral reasons and the context in which they use them to deliberate.

Thirdly, recall the Weak Restraints Objection I made against the Restrained Domination Conception. The crux of that objection was that explaining the legitimacy of coercively enforced political principles by the fact that it is the output of democratic majoritarian decision-making offered no real restraint on the sort of political order the majority could create. They could dominate the minority which tries to change the political order by democratic means, or move away from majoritarianism as the basis of the political order itself. As such, the conception cannot show how reasonable people can plausibly maintain a political order. However, the Dual Convergent Conception avoids this objection by committing to stronger constraints on the use of coercive power. It is committed to the impermissibility of coercively enforcing any political principle that, is not selected by the procedure prescribed by the political norm or, is not within the set of political principles that all have sufficient reason to accept. This means that the political order is more easily maintained given it can only be changed by a procedure prescribed by a converged upon political norm, and only changed in accordance with political principles from a set that has been converged upon. The dual convergences provide a certain level of rigidity to the political order.

3 Modus Vivendi Objection

One objection political liberals might have is that the Dual Convergent Conception is nothing but an elaborate description of a modus vivendi.[5] That is to say, the idea of "ordered moral warfare" in which reasonable people have sufficient moral reason to *accept* a political principle involves a commitment to coordinate based on a contingent balance of forces. This is because as soon as a group of reasonable people have the means to take power and realise the political principles they endorse, they will. After all they do not coordinate because they have sufficient moral reason to endorse a political principle understood as internalising a political principle as one's own. Rather they coordinate because they have sufficient moral reason to accept it which is understood as a purely practical attitude of freely complying with a political principle. As such, when they need not accept a political principle they have no reason to comply. But, so the objection goes, this is not the sort of stability reasonable people should strive for. This is a political order predicated entirely on the contingent fact that no group of reasonable has enough resources to oppress their fellow citizens. It is not stable because people have reason to endorse it, but rather purely for the instrumental reason that one cannot successfully dominate those they disagree with.

This objection has some merit when put against political realists. Some have flirted with the idea of cashing out political legitimacy with a modus vivendi where people converge for purely reasons of self-interest.[6] However, in the case of the Dual Convergent Conception, the objection misreads what the idea of "ordered moral warfare" is. The convergent agreements that constitute it do not involve people converging for reasons related to not being able to overpower their opponents. Rather reasonable people converge based on the considerations provided by their concept of JUSTICE, the social context they occupy, and the need to achieve the good of political legitimacy. This means they recognise given their own concept of justice, and the fact of reasonable disagreement about justice that achieving the good of political legitimacy requires reconciliation in the long term. Without it there is no real possibility that, they and those they care about can develop and maintain their capacities for a conception of justice, and in turn realise the political principles they endorse.

The objection also ignores the point that one of the agreements that constitute "ordered moral warfare" involves a convergence on a political norm. This agreement

[5] See Rawls (2005: 146–149) for the canonical political liberal view of a modus vivendi and its problems.

[6] See Horton (2010: 437–442). See also Sleat (2013: Ch. 4) for an overview of this move by some political realists.

involves reasonable people accepting a set of procedures or social practices that requires them, at some stage, to relinquish power and have society transition to enforcing different political principles. In short, the convergence on the political norm resists the temptation to impose the principles that one's concept of JUSTICE supports as soon as one has the power to do so.

4 Conceptual Integrity Objection

Another objection political liberals might have is that, one of my objections to the Convergence Conception of political liberalism in Chapter 4, applies equally to the Dual Convergent Conception. Recall, I objected that the Social Equilibrium version of the Convergence Conception faces the Conceptual Integrity Objection because it threatens people's conceptual integrity. I argued that it curbs people's autonomy to use the concepts they believe represent reality accurately. But, in proposing the Dual Convergent Conception I adopted a Social Equilibrium model for how reasonable people can converge on a political norm. It seems political liberals could then respond that this will involve violating people's conceptual integrity as much as I claimed it would in the case of modelling reasonable people converging on social norms. If I deny that people's conceptual integrity is not violated in converging on a political norm I have no grounds for objecting that it will in the case of social norms and so the Dual Convergent Conception and the political liberal Convergence Conception are on equal footing.

This objection, however, ignores the nature of the political realist normative standard the Dual Convergent Conception uses. Specifically, it ignores how in the political realist's normative standard of "meeting the BLD", the attitude elicited by conclusive justification is acceptance, and not endorsement. This means that on the Dual Convergent Conception, the convergence on the political norm is an agreement on *accepting* a political norm. As such it does not threaten a person's conceptual integrity because they are not having to endorse a political norm for reasons of social pressure. They do not have to internalise the political norm as their own ideal political norm. Rather they merely have to freely commit to it as a political norm to comply with. As such they will have sufficient moral reason to accept a political norm, but at the same time endorse their ideal political norm and advocate for their society to converge on it.

5 Conclusion

I started this thesis with what Rawls (1999: 514) says in the following passage about the sort of perspective that underwrites the normative force of his arguments:

> The perspective of eternity is not a perspective from a certain place beyond the world, nor the point of view of a transcendent being; rather it is a certain form of thought and feeling that rational persons can adopt within the world. And having done so, they can, whatever their generation, bring together into one scheme all individual perspectives and arrive together at regulative principles that can be affirmed by everyone as he lives by them, each from his own standpoint. Purity of heart, if one could attain it, would be to see clearly and to act with grace and self-command from this point of view.

I believe we have now reached where Rawls hoped to. We have arrived at a theory, or at least a sketch of a theory, that tells us how to act from the perspective of eternity. That is, from the perspective of reasonable disagreement about justice itself. The Dual Convergent Conception I have argued for prescribes how reasonable people, when they take seriously the reasonable disagreement about justice between themselves, ought to go about achieving a stable political order.

I defended this theory on the basis of a Disagreement to Legitimacy argument. This involved first finding the best explanation of reasonable disagreement about justice. After arguing against extant explanations I proposed Diverse Packages Theory as the best explanation of reasonable disagreement. This is the theory that says what best explains why reasonable people make conflicting judgements about the institutions and outcomes that justice requires, is that reasonable people possess and use diverse concepts *and* conceptions of justice.

I then proceeded to find the theory of political legitimacy that, given Diverse Packages Theory, can show how reasonable people could achieve a stable political order. That is, a theory that can show how reasonable people's balance of reasons can provide them and continue to provide them sufficient moral reason to coordinate on coercively enforced political principles or rules. After arguing against extant conceptions of political liberalism and political realism I proposed the Dual Convergent Conception of political realism as the theory of political legitimacy that can achieve that. This theory combines elements of the Social Equilibrium view of convergence with the political realist normative standard for justification into the central organising idea of ordered moral warfare. By doing this the theory shows how reasonable

people who disagree so deeply that they possess divergent concepts of justice can live together in a stable political order. In short, how they ought to act from the perspective of eternity.

Bibliography

Adams, David. "Knowing when Disagreements are Deep". In: *Informal Logic* 25.1 (1985), pp. 65–77.

Ball, Derek. "Revisionary Analysis without Meaning Change (Or, Could Women Be Analytically Oppressed?)" In: *Conceptual Engineering and Conceptual Ethics*. Ed. by Alexis Burgess, Herman Cappelen, and David Plunkett. Oxford University Press, 2020.

Ballantyne, Nathan. "Verbal Disagreements and Philosophical Scepticism". In: *Australasian Journal of Philosophy* 94.4 (2016), pp. 752–765.

Barnidge, Matthew. "Exposure to Political Disagreement in Social Media Versus Face-to-Face and Anonymous Online Settings". In: *Political Communication* 34.2 (2017), pp. 302–321.

Bellamy, Richard. *Political Constitutionalism: A Republican Defence of the Constitutionality of Democracy*. Cambridge University Press, 2007.

Besch, Thomas. "Political Liberalism, the Internal Conception, and the Problem of Public Dogma". In: *Philosophy and Public Issues* 2.1 (2012), pp. 153–177.

Bicchieri, Cristina. *Norms in the Wild: How to Diagnose, Measure and Change Social Norms*. Cambridge University Press, 2016.

— *The Grammar of Society: The Nature and Dynamics of Norms*. Cambridge University Press, 2006.

Bicchieri, Cristina and Peter McNally. "Shrieking Sirens: Schemta, Scripts, and Social Norms. How Change Occurs". In: *Social Philosophy and Policy* 35.1 (2018), pp. 23–53.

Bird, Colin. "Coercion and public justification". In: *Politics, Philosophy & Economics* 13.3 (2014), pp. 189–214.

Bloom, Paul and Karen Wynn. "What Develops in Moral Development". In: *Core Knowledge and Conceptual Change*. Ed. by David Barner and Andrew Scott Baron. Oxford University Press, 2016.

Boettcher, James W. "Against the Asymmetric Convergence Model of Public Justification". In: *Ethical Theory and Moral Practice* 18.1 (2015), pp. 191–208.

Brink, David. *Moral Realism and the Foundations of Ethics*. Cambridge University Press, 1989.

Buchanan, Allen. "Political Legitimacy and Democracy". In: *Ethics* 112.4 (2002), pp. 689–719.

Burgess, Alexis and David Plunkett. "Conceptual Ethics I". In: *Philosophy Compass* 8.12 (2013), pp. 1091–1101.

— "Conceptual Ethics II". In: *Philosophy Compass* 8.12 (2013), pp. 1102–1110.

Cacioppo, John T. and William Patrick. *Loneliness: Human Nature And The Need For Social Connection*. W. W Norton & Company, 2009.

Campos, Paul. "Secular Fundamentalism". In: *Columbia Law Review* 94.6 (1994), pp. 1814–1827.

Cappelen, Herman. *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press, 2018.

Cappelen, Herman and David Plunkett. "Introduction". In: *Conceptual Engineering and Conceptual Ethics*. Ed. by Alexis Burgess, Herman Cappelen, and David Plunkett. Oxford University Press, 2020.

Carey, Brian. "Public Reason – Honesty, Not Sincerity". In: *The Journal of Political Philosophy* 26.1 (2018), pp. 47–64.

Carey, Susan. "Précis of The Origin of Concepts". In: *Behavioral and Brain Sciences* 34.3 (2011), pp. 113–124.

— *The Origin of Concepts*. Oxford University Press, 2009.

— "Why Theories of Concepts Should Not Ignore the Problem of Acquisition". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. Oxford University Press, 2015.

Chalmers, David. "Verbal Disputes". In: *Philosophical Review* 11.4 (2011), pp. 515–566.

Christiano, Thomas. *The Constitution of Equality: Democratic Authority and its Limits*. Oxford University Press, 2009.

Chung, Hun. "The Impossibility of Liberal Rights in a Diverse World". In: *Economics and Philosophy* 35.1 (2019), pp. 1–27.

Chung, Hun and Brian Kogelmann. "Diversity and rights: a social choice-theoretic analysis of the possibility of public reason". In: *Synthese* 197 (2020), pp. 839–865.

Cohen, Gerald. *Rescuing Justice and Equality*. Cambridge, Mass.: Harvard University Press, 2008.

D'Agostino, Fred. *Free Public Reason: Making It Up As We Go*. Oxford University Press, 1996.

Dreben, Burton. "On Rawls and Political Liberalism". In: *The Cambridge Companion to Rawls*. Ed. by Samuel Freeman. Cambridge University Press, 2003.

Driver, Julia. "The Limits of the Dual-Process View". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.

Dworkin, Ronald. *Justice for Hedgehogs*. Harvard University Press, 2011.

Eberle, Christopher. "Consensus, Convergence, and Religiously Justified Coercion". In: *Public Affairs Quarterly* 25.4 (2011), pp. 281–303.

— *Religious Conviction in Liberal Politics*. Cambridge University Press, 2002.

Edmundson, William A. *Three Anarchical Fallacies*. Cambridge University Press, 1998.

Elster, Jon. "Political Norms". In: *Iyyun: The Jerusalem Philosophical Quarterly* 63 (2014), pp. 47–59.

Enoch, David. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press, 2011.

Erman, Eva and Niklas Moller. "Political Legitimacy for Our World: Where Is Political Realism Going?" In: *The Journal of Politics* 80.2 (2018), pp. 525–538.

— "Political Legitimacy in the Real Normative World: The Priority of Morality and the Autonomy of the Political". In: *British Journal of Political Science* 45.1 (2013), pp. 215–233.

— "Practices and Principles: On the Methodological Turn in Political Theory". In: *Philosophy Compass* 10.8 (2015), pp. 533–546.

— "Three Failed Charges Against Ideal Theory". In: *Social Theory and Practice* 39.1 (2013), pp. 19–44.

— "What distinguishes the practice-dependent approach to justice?" In: *Philosophy and Social Criticism* 42.1 (2016), pp. 3–23.

— "Why Political Realists Should Not Be Afraid of Moral Values". In: *Journal of Philosophical Research* 40.4 (2015), pp. 459–464.

Estlund, David. "Methodological moralism in political philosophy". In: *Critical Review of International Social and Political Philosophy* 20.3 (2017), pp. 385–402.

Evans, Isobel E. M. et al. "Social isolation, cognitive reserve, and cognition in healthy older people". In: *PLOS One* 13.8 (2018), pp. 1–14.

Finlayson, Lorna. "With radicals like these, who needs conservatives? Doom, gloom, and realism in political theory". In: *European Journal of Political Theory* 16.3 (2017), pp. 264–282.

Fogelin, Robert. "The logic of deep disagreements". In: *Informal Logic* 7.1 (1985), pp. 1–8.

Forrester, Katrina. "Judith Shklar, Bernard Williams and political realism". In: *European Journal of Political Theory* 11.3 (2012), pp. 247–272.

Fowler, Timothy and Zofia Stemplowska. "The Asymmetry Objection Rides Again: On the Nature and Significance of Justificatory Disagreement". In: *Journal of Applied Philosophy* 32.2 (2015), pp. 133–146.

Fraassen, Bas C. van. *The Scientific Image*. Oxford University Press, 1980.

Frances, Bryan. *Disagreement*. Polity Press, 2014.

Freyenhagen, Fabian. "Taking reasonable pluralism seriously: an internal critique of political liberalism". In: *Politics, Philosophy & Economics* 10.3 (2011), pp. 323–342.

Galston, William A. "Realism in political theory". In: *European Journal of Political Theory* 9.4 (2010), pp. 385–411.

Gaus, Gerald. "A Tale of Two Sets: Public Reason in Equilibrium". In: *Public Affairs Quarterly* 25.4 (2011), pp. 305–325.

— "Is Public Reason a Normalization Project? Deep Diversity and the Open Society". In: *Social Philosophy Today* 33.1 (2017), pp. 27–52.

— "Reasonable Pluralism and the Domain of the Political: How the Weaknesses of John Rawls's Political Liberalism Can be Overcome by a Justificatory Liberalism". In: *Inquiry* 42.2 (1999), pp. 259–284.

— "The Complexity of a Diverse Moral Order". In: *The Georgetown Journal of Law & Public Policy* 16.1 (2018), pp. 645–680.

— "The Diversity of Comprehensive Liberalisms". In: *Handbook of Political Theory*. Ed. by Gerald F. Gaus and Chandran Kukathas. SAGE Publications, 2004.

— *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press, 2011.

— "The Turn to a Political Liberalism". In: *A Companion to Rawls*. Ed. by Jon Mandle and David A. Reidy. Wiley Blackwell, 2014.

— *The Tyranny of the Ideal: Justice in a Diverse society*. Princeton university Press, 2016.

Gaus, Gerald and Kevin Vallier. "The roles of religious conviction in a publicly justified polity: The implications of convergence, asymmetry and political institutions". In: *Philosophy Social Criticism* 35.1–2 (2009), pp. 51–76.

Geuss, Raymond. *Philosophy and Real Politics*. Princeton University Press, 2008.

Gibbard, Allan. *Thinking How to Live*. Harvard University Press, 2003.

Green, Leslie. *The Authority of the State*. Oxford University Press, 1988.

Greene, Joshua. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics". In: *Ethics* 124.4 (2014), pp. 695–726.

— *Moral Tribes: Emotion, Reason, And The Gap Between Us And Them*. The Penguin Press, 2013.
— "Reply to Driver and Darwall". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.
— "The Cognitive Neuroscience of Moral Judgment and Decision Making". In: *The Cognitive Neurosciences*. Ed. by Michael S. Gazzaniga and George R. Mangun. MIT Press, 2014.
Gutting, Gary. *Religious Belief and Religious Skepticism*. Notre Dame: University of Notre Dame Press, 1982.
Haidt, Jonathan. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. The Penguin Press, 2012.
Hall, Edward. "Bernard Williams and the Basic Legitimation Demand: A Defence". In: *Political Studies* 63.2 (2015), pp. 466–480.
— "How to do realistic political theory (and why you might want to)". In: *European Journal of Political Theory* 16.3 (2017), pp. 283–303.
Hamlin, J. Kiley. "Does the Infant Possess a Moral Concept?" In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. Oxford University Press, 2015.
— "The infantile origins of our moral brains". In: *The moral brain: A multidisciplinary perspective*. Ed. by J. Decety and T. Wheatley. MIT Press, 2017.
Hampton, James A. "Concepts in the Semantic Triangle". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. Oxford University Press, 2015.
Hare, R. M. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Clarendon Press, 1981.
Hartley, Christie and Lori Watson. *Equal Citizenship and Public Reason: A Feminist Political Liberalism*. Oxford University Press, 2018.
— "Feminism, Religion, And Shared Reasons: A Defense Of Exclusive Public Reason". In: *Law and Philosophy* 28.5 (2009), pp. 493–536.
Hazlett, Allan. "Entitlement and Mutually Recognized Reasonable Disagreement". In: *Episteme* 11.1 (2014), pp. 1–25.
Horton, John. "Realism, liberal moralism and a political theory of modus vivendi". In: *European Journal of Political Theory* 9.4 (2010), pp. 431–448.
Huckfeldt, Robert, Paul E. Johnson, and John Sprague. *Political Disagreement: The Survival Of Diverse Opinions Within Communication Networks*. Cambridge University Press, 2004.

Huckfeldt, Robert and Jeanette Morehouse Mendez. "Moths, Flames, and Political Engagement: Managing Disagreement within Communication Networks". In: *The Journal of Politics,* 70.1 (2008), pp. 83–96.

Jacobson, Daniel. "Moral Dumbfounding and Moral Stupefaction". In: *Oxford Studies in Normative Ethics: Volume 2*. Ed. by Mark Timmons. Oxford University Press, 2012.

Jubb, Robert. "On What a Distinctively Political Normativity Is". In: *Political Studies Review* 17.4 (2019), pp. 360–369.

— "Playing Kant at the Court of King Arthur". In: *Political Studies* 63.4 (2015), pp. 919–934.

Jubb, Robert and Enzo Rossi. "Political Norms and Moral Values". In: *Journal of Philosophical Research* 40.4 (2015), pp. 445–458.

— "Why Moralists Should Be Afraid of Political Values: A Rejoinder". In: *Journal of Philosophical Research* 40.4 (2015), pp. 465–468.

Kahane, Guy. "Is, Ought, and the Brain". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.

Kahane, Guy et al. "The neural basis of intuitive and counterintuitive moral judgment". In: *Social Cognitive and Affective Neuroscience* 7.4 (2012), pp. 393–402.

Kalish, Charles W. "Normative Concepts". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. Oxford University Press, 2015.

Kappel, Klemens. "Higher Order Evidence and Deep Disagreement". In: *Topoi* (2018), pp. 1–12. URL: https://doi.org/10.1007/s11245-018-9587-8.

Kelly, Erin and Lionel McPherson. "On Tolerating the Unreasonable". In: *The Journal of Political Philosophy* 9.1 (2001), pp. 38–55.

Kelly, Thomas. "The Epistemic Significance of Disagreement". In: *Oxford Studies In Epistemology: Volume 1*. Ed. by Tamar Szabó Gendler and John Hawthorne. Oxford University Press, 2005.

Kennett, Jeanette and Cordelia Fine. "Will the Real Moral Judgment Please Stand Up?: The Implications of Social Intuitionist Models of Cognition for Metaethics and Moral Psychology". In: *Ethical Theory and Moral Practice* 12.1 (2009), pp. 77–96.

Kennett, Jeanette and Philip Gerrans. "The Rationalist Delusion?: A Post Hoc Investigation". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.

King, Nathan. "Disagreement: What's the Problem? or A Good Peer is Hard to Find". In: *Philosophy and Phenomenological Research* 85.2 (2012), pp. 249–272.

Knight, Carl. "Justice for Foxes". In: *Law and Philosophy* 34.6 (2006), pp. 633–659.

Kogelmann, Brian. "Justice, Diversity, and the Well-Ordered Society". In: *The Philosophical Quarterly* 67.269 (2017), pp. 663–684.

Kohlberg, Lawrence. *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. San Francisco: Harper and Row, 1984.

Kumar, Victor and Joshua May. "On Rawls and Political Liberalism". In: *Methodology and Moral Philosophy*. Ed. by Jussi Suikkanen and Antti Kauppinen. Routledge, 2018.

Kölbel, Max. "Faultless Disagreement". In: *Proceedings of the Aristotelian Society* 104.1 (2004), pp. 53–73.

LaFollette, Hugh and Michael L. Woodruff. "Reflection and Reasoning in Moral Judgment". In: *Philosophical Psychology* 28.3 (2015), pp. 452–465.

Larmore, Charles. "Political Liberalism". In: *Political Theory* 18.3 (1990), pp. 339–360.

— "The Moral Basis Of Political Liberalism". In: *The Journal of Philosophy* 96.12 (1999), pp. 599–625.

— "What Is Political Philosophy?" In: *Journal of Moral Philosophy* 10.3 (2013), pp. 276–306.

Leland, R. J. and Han van Wietmarschen. "Political Liberalism and Political Community". In: *Journal of Moral Philosophy* 14.2 (2017), pp. 142–167.

— "Reasonableness, Intellectual Modesty, and Reciprocity in Political Justification". In: *Ethics* 122.4 (2012), pp. 721–747.

Lister, Andrew. *Public Reason and Political Community*. Bloomsbury, 2013.

Locke, John. *An Essay Concerning Human Understanding*. Oxford University Press, 2008.

Macedo, Stephen. *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism*. Oxford University Press, 1991.

Machery, Edouard. "By Default: Concepts Are Accessed in a Context-Independent Manner". In: *The Conceptual Mind: New Directions in the Study of Concepts*. Ed. by Eric Margolis and Stephen Laurence. Oxford University Press, 2015.

— *Doing without Concepts*. Oxford University Press, 2009.

— "Précis of Doing without Concepts". In: *Behavioral and Brain Sciences* 33.2–3 (2010), pp. 195–206.

Margolis, Eric and Stephen Laurence. "Concepts and Cognitive Science". In: *Concepts: Core Readings*. Ed. by Eric Margolis and Stephen Laurence. Bradford Books, 1999.

Mason, Andrew. *Explaining Political Disagreement*. Cambridge University Press, 1993.

— "Rawlsian Theory and the Circumstances of Politics". In: *Political Theory* 38.5 (2010), pp. 658–683.

Matheson, Jonathan. "Disagreement: Idealized and Everyday". In: *The Ethics of Belief: Individual and Social*. Ed. by Jonathan Matheson and Rico Vitz. Oxford University Press, 2014.

Maynard, Jonathan Leader and Alex Worsnip. "Is There a Distinctively Political Normativity?" In: *Ethics* 128.4 (2018), pp. 756–787.

McClurg, Scott D. "Political Disagreement in Context: The Conditional Effect of Neighborhood Context, Disagreement and Political Talk on Electoral Participation". In: *Political Behaviour* 28.4 (2006), pp. 349–366.

McDowell, John. *Mind, Value, and Reality*. Harvard University Press, 1998.

McMahon, Christopher. *Reasonable Disagreement: A Theory of Political Morality*. Cambridge University Press, 2009.

— *Reasonableness and Fairness: A Historical Theory*. Cambridge University Press, 2016.

McQueen, Alison. "The Case for Kinship: Classical Realism and Political Realism". In: *Politics Recovered: Essays on Realist Political Thought*. Ed. by Matt Sleat. Columbia University Press, 2018.

Midgley, Mary. "Philosophical Plumbing". In: *Royal Institute of Philosophy Supplement* 30 (1992), pp. 139–151.

Miller, David. *Justice for Earthlings: Essays in Political Philosophy*. Cambridge University Press, 2013.

Muldoon, Ryan. *Social Contract Theory for a Diverse World: Beyond Tolerance*. Routledge, 2016.

Mutz, Diana C. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge University Press, 2006.

Nagel, Thomas. *Equality and Partiality*. Oxford University Press, 1991.

— "Moral Conflict and Political Legitimacy". In: *Philosophy & Public Affairs* 16.3 (1987), pp. 215–240.

Newey, Glen. "Two dogmas of liberalism". In: *European Journal of Political Theory* 9.4 (2010), pp. 449–465.

Nichols, Shaun. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press, 2004.

Nir, Lilach. "Disagreement and Opposition in Social Networks: Does Disagreement Discourage Turnout?" In: *Political Studies* 674–692.3 (2011), pp. 149–160.

Nucci, Larry, Elliot Turiel, and Alona D. Roded. "Continuities and Discontinuities in the Development of Moral Judgments". In: *Human Development* 60.6 (2017), pp. 279–341.

Nussbaum, Martha C. "Perfectionist Liberalism and Political Liberalism". In: *Philosophy & Public Affairs* 39.1 (2011), pp. 3–45.

Page, Scott E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.

Patterson, Richard, Jared Rothstein, and Aron K. Barbey. "Reasoning, cognitive control, and moral intuition". In: *Frontiers in Integrative Neuroscience* 6.114 (2012), pp. 1–8.

Paxton, Joseph M., Leo Ungar, and Joshua Greene. "Reflection and Reasoning in Moral Judgment". In: *Cognitive Science* 36.1 (2011), pp. 1–15.

Peter, Fabienne. "Epistemic Foundations of Political Liberalism". In: *Journal of Moral Philosophy* 10.5 (2013), pp. 598–620.

Philip, Mark. "Realism without Illusions". In: *Political Theory* 40.5 (2012), pp. 629–649.

— "What is to be done? Political theory and political realism". In: *European Journal of Political Theory* 9.4 (2010), pp. 466–484.

Piaget, Jean. *The Moral Judgment of the Child*. New York: Free Press, 1965.

Plunkett, David. "Which Concepts Should We Use?: Metalinguistic Negotiations and The Methodology of Philosophy". In: *Inquiry* 58.7–8 (2015), pp. 828–874.

Plunkett, David and Timothy Sundell. "Disagreement and the Semantics of Normative and Evaluative Terms". In: *Philosophers' Imprint* 13.23 (2013), pp. 1–37.

— "Dworkin's Interpretivism And The Pragmatics Of Legal Disputes". In: *Legal Theory* 19.3 (2013), pp. 242–281.

Prinz, Jesse. "Sentimentalism and the Moral Brain". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.

Pritchard, Duncan. "Wittgensteinian Hinge Epistemology and Deep Disagreement". In: *Topoi* (2018), pp. 1–9. URL: https://doi.org/10.1007/s11245-018-9612-y.

Quong, Jonathan. *Liberalism Without Perfection*. Oxford University Press, 2011.

Railton, Peter. "Moral Learning: Conceptual foundations and normative relevance". In: *Cognition* 167 (2017), pp. 172–190.

Ranalli, Chris. "Deep disagreement and hinge epistemology". In: *Synthese* (2018), pp. 1–33. URL: https://doi.org/10.1007/s11229-018-01956-2.

— "What is Deep Disagreement?" In: *Topoi* (2018), pp. 1–16. URL: https://doi.org/10.1007/s11245-018-9600-2.

Rawls, John. *A Theory of Justice: Revised Edition*. Cambridge, Mass.: Belknap Press of Harvard University Press, 1999.

— *Political Liberalism: Expanded Edition*. New York: Columbia University Press, 2005.

Raz, Joseph. *The Morality of Freedom*. Oxford University Press, 1986.

— "The Problem of Authority: Revisiting the Service Conception". In: *Minnesota Law Review* 90.4 (2006), pp. 1003–1044.

Reidy, David A. "Rawls's Wide View Of Public Reason: Not Wide Enough". In: *Res Publica* 6.1 (2000), pp. 49–72.

— "Reciprocity And Reasonable Disagreement: From Liberal To Democratic Legitimacy". In: *Philosophical Studies* 132.2 (2007), pp. 243–291.

Reilly, Jack Lyons. "Social connectedness and political behavior". In: *Research & Politics* 4.3 (2017), pp. 1–8.

Rey, Georges. "Concepts and conceptions: A reply to Smith, Medin and Rips". In: *Cognition* 19.3 (1985), pp. 297–303.

— "Concepts and stereotypes". In: *Cognition* 15.1–3 (1983), pp. 237–262.

Ripstein, Arthur. "Authority and Coercion". In: *Philosophy & Public Affairs* 32.1 (2004), pp. 2–35.

Roberts, Craige. "Information structure in discourse: Towards an integrated formal theory of pragmatics". In: *Semantics & Pragmatics* 5.1 (2012), pp. 1–69.

Rossi, Enzo. "Being realistic and demanding the impossible". In: *Constellations* 26.4 (2019), pp. 638–652.

— "Consensus, compromise, justice and legitimacy". In: *Critical Review of International Social and Political Philosophy* 16.4 (2013), pp. 557–572.

— "Justice, legitimacy and (normative) authority for political realists". In: *Critical Review of International Social and Political Philosophy* 15.2 (2012), pp. 149–164.

Rossi, Enzo and Matt Sleat. "Realism in Normative Political Theory". In: *Philosophy Compass* 9.10 (2014), pp. 689–701.

Sangiovanni, Andrea. "Justice and the Priority of Politics to Morality". In: *The Journal of Political Philosophy* 16.2 (2008), pp. 137–164.

Sauer, Hanno. "Can't We All Disagree More Constructively? Moral Foundations, Moral Reasoning, and Political Disagreement". In: *Neuroethics* 8.2 (2015), pp. 153–169.

— "Educated intuitions. Automaticity and rationality in moral judgement". In: *Philosophical Explorations* 15.3 (2012), pp. 255–275.

Sawyer, Sarah. "Subjective Externalism". In: *Theoria* 84.1 (2018), pp. 4–22.

— "Talk and Thought". In: *Conceptual Engineering and Conceptual Ethics*. Ed. by Alexis Burgess, Herman Cappelen, and David Plunkett. Oxford University Press, 2020.

— "The Importance of Concepts". In: *Proceedings of the Aristotelian Society* 118.2 (2018), pp. 1–21.

Schafer-Landau, Russ. *Moral Realism: A Defence*. Oxford University Press, 2003.

Schoelandt, Chad van. "Rawlsian Functionalism and the Problem of Coordination". In: *Social Theory and Practice* forthcoming (2020).

Schoelandt, Chad Van. "Justification, coercion, and the place of public reason". In: *Philosophical Studies* 172.4 (2015), pp. 1031–1050.

Schroeter, Laura and François Schroeter. "Normative Concepts: A Connectedness Model". In: *Philosophers' Imprint* 14.25 (2014), pp. 1–26.

Schwartzman, Micah. "The completeness of public reason". In: *Politics, Philosophy & Economics* 3.2 (2004), pp. 191–220.

Schwartzman, Micah and Jocelyn Wilson. "The Unreasonableness of Catholic Integralism". In: *San Diego Law Review* 56.4 (2019), pp. 1039–1068.

Sen, Amartya. *The Idea of Justice*. Penguin Books, 2010.

Sher, George. "Moral Thinking: Its Levels, Method, and Point. by R. M. Hare". In: *Nous* 18.1 (1984), pp. 179–184.

Shklar, Judith. "The Liberalism of Fear". In: *Liberalism and the Moral Life*. Ed. by Nancy L. Rosenblum. Harvard University Press, 1989.

Shotter, Andrew and Barry Sopher. "Social Learning and Coordination Conventions in Intergenerational Games: An Experimental Study". In: *Journal of Political Economy* 113.3 (2003), pp. 498–529.

Simmons, John. *Justification and Legitimacy: Essays on Rights and Obligations*. Cambridge University Press, 2001.

Sleat, Matt. "Bernard Williams and the possibility of a realist political theory". In: *European Journal of Political Theory* 9.4 (2010), pp. 485–503.

— *Liberal realism: A realist theory of liberal politics*. Manchester University Press, 2013.

Stanley, Matthew, Siyuan Yin, and Walter Sinnott-Armstrong. "A reason-based explanation for moral dumbfounding". In: *Judgement and Decision Making* 14.2 (2019), pp. 120–129.

Sudarshan, Saranga. "The Independence of Political Theory". In: *manuscript* ().

Suhler, Christopher L. and Patricia Churchland. "Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory". In: *Journal of Cognitive Neuroscience* 23.9 (2011), pp. 2103–2116.

Thrasher, John and Kevin Vallier. "Political Stability in the Open Society". In: *American Journal of Political Science* 62.2 (2018), pp. 398–409.

Turiel, Elliot. *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge University Press, 2002.

Valentini, Laura. "Ideal vs. Non-ideal Theory: A Conceptual Map". In: *Philosophy Compass* 7.9 (2012), pp. 654–664.

— "Justice, Disagreement and Democracy". In: *British Journal of Political Science* 43.1 (2013), pp. 177–199.

Vallier, Kevin. "Convergence and Consensus in Public Reason". In: *Public Affairs Quarterly* 25.4 (2011), pp. 261–280.

— *Liberal Politics and Public Faith: Beyond Separation*. Routledge, 2014.

— "Liberalism, Religion And Integrity". In: *Australasian Journal of Philosophy* 90.1 (2012), pp. 149–165.

— *Must Politics Be War?: Restoring Our Trust in the Open Society*. Oxford University Press, 2019.

Vanderschraaf and Brian Skyrms. "Learning to Take Turns". In: *Erkenntnis* 59 (2003), pp. 311–347.

Waldron, Jeremy. "Isolating Public Reasons". In: *Rawls's Political Liberalism*. Ed. by Thom Brooks and Martha C. Nussbaum. Columbia University Press, 2015.

— *Law and Disagreement*. Oxford University Press, 1999.

Weithman, Paul. "Legitimacy and the Project of Political Liberalism". In: *Rawls's Political Liberalism*. Ed. by Thom Brooks and Martha C. Nussbaum. Columbia University Press, 2015.

— *Why Political Liberalism?: On John Rawls's Political Turn*. Oxford University Press, 2010.

Wenar, Leif. "Political Liberalism: An Internal Critique". In: *Ethics* 106.1 (1995), pp. 32–62.

Wietmarschen, Han van. "Reasonable Citizens and Epistemic Peers: A Skeptical Problem for Political Liberalism". In: *The Journal of Political Philosophy* 26.4 (2018), pp. 486–507.

Wiggins, David. *Ethics: Twelve Lectures on the Philosophy of Morality*. Harvard University Press, 2006.

Williams, Andrew. "The Alleged Incompleteness of Public Reason". In: *Res Publica* 6.2 (2000), pp. 199–211.

Williams, Bernard. *In the Beginning Was the Deed: Realism And Moralism In Political Argument*. Princeton University Press, 2005.

— *Truth & Truthfulness: An Essay in Genealogy*. Princeton University Press, 2002.

Winter, Jack. "Justice for Hedgehogs, Conceptual Authenticity for Foxes: Ronald Dworkin on Value Conflicts". In: *Res Publica* 22.4 (2016), pp. 463–479.

Wojcieszak, Magdalena E. "Pulling Toward or Pulling Away:Deliberation, Disagreement, and Opinion Extremity in Political Participation". In: *Social Science Quarterly* 92.1 (2011), pp. 206–225.

Wojcieszak, Magdalena E. and Vincent Price. "Perceived Versus Actual Disagreement: Which Influences Deliberative Experiences?" In: *Journal of Communication* 62.3 (2012), pp. 418–436.

Wolterstorff, Nicholas. "The Role of Religion in Decision and Discussion of Political Issues". In: *Religion in the Public Square: The Place of Religious Convictions in Political Debate*. Ed. by James P. Sterba and Rosemarie Tong. Rowman & Littlefield, 1997.

Woodward, James. "Emotion versus Cognition in Moral Decision-Making: A Dubious Dichotomy". In: *Moral Brains: The Neuroscience Of Morality*. Ed. by Matthew Liao. Oxford University Press, 2016.

— *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

Young, Liane and James Dungan. "Where in the brain is morality? Everywhere and maybe nowhere". In: *Social Neuroscience* 7.1 (2012), pp. 1–10.