Political Contractarianism

Justice as Mutual Advantage

Ian J. Carroll
Nuffield College, Oxford

This thesis is submitted in partial fulfilment of the requirements of the degree of DPhil in Politics in the Department of Politics and International Relations at the University of Oxford.

Title	Political Contractarianism: Justice as Mutual Advantage
Author	Ian J. Carroll, Nuffield College
Degree	DPhil Politics
Submitted	Michaelmas, 2018 (Corrected Trinity, 2019)

Abstract

Justice is the highest of our moral ends. Implementing a conception of justice requires the coordination of many people. However, contemporary societies are characterised by pervasive and persistent disagreement about justice, and individuals in such societies typically lack the power to simply impose a conception of justice. This presents a problem: How can one best pursue the implementation of justice when (a) one lacks the power to impose it unilaterally and (b) others cannot be persuaded that one's conception of justice is correct?

In this context, I propose that citizens implement a conception of justice which represents a rational compromise (a *modus vivendi*) between all of the particular conceptions of justice held by those citizens whose coordination is necessary to implement the compromise. I model the 'pure' conceptions of justice endorsed by citizens as competing preference orderings. I then use a contractarian bargaining theory to identify a point of compromise at which each citizen has decisive reason to accept and implement the proposed bargain.

I argue that, given certain conditions, the political institutions which emerge from the bargaining situation are robustly liberal. This result is (within broad limits) independent of the distribution of pure conceptions of justice among the citizens. That is, liberalism follows simply from the circumstances of disagreement, rather than from commitments already present among citizens.

My conclusion is novel in two further respects: First, unlike earlier *modus vivendi* accounts, I do not rely on any minimal moral premises (thus avoiding standard objections to such accounts). Second, unlike earlier contractarian theories, I argue that the political institutions identified by my model will be strongly egalitarian. I do this by developing a new understanding of the bargaining dynamics operating in the hypothetical choice situation – the side payments approach – and hence a new basis of equality for contractarian theories.

Acknowledgements

Even for a doctoral thesis, this project was a long time in the making. Along the way, I have incurred more debts than I can acknowledge properly here. First and foremost, I am grateful to my supervisors, Dan McDermott and David Miller, for their insightful comments, advice, and criticism; for their assistance in navigating the process of writing and submitting the thesis; and for exemplifying the philosophical clarity and rigour to which students like me aspire. Although they entered rather later in the process, I also owe a significant debt to my examiners, Zosia Stemplowska and Jerry Gaus. Without their questions and objections, my argument would undeniably have been poorer. By the time I came to write these acknowledgements, Professor Gaus had sadly passed away. I regret that we didn't have more time to argue about these problems, but I am very thankful for the generous advice and assistance he gave to me.

For helpful comments on earlier drafts of the work presented here, I am grateful to Alice Baderin, Dan Butt, Jerry Cohen, Faik Kurtulmuş, Seth Lazar, Chris Nathan, Stuart White, and Alexa Zellentin; and to audiences at Oxford's Graduate Political Theory Workshop, at the Nuffield Political Theory Workshop, at the *Liberal Neutrality: A Re-evaluation* conference organised by le Centre de recherche en éthique de l'Université de Montréal at McGill University, and at the *Contractarian Moral Theory: the 25th Anniversary of* Morals By Agreement conference organised at York University, Ontario. While working on this project, I was supported by a studentship from the Arts and Humanities Research Council. Some of the work was also conducted while I was a lecturer at Brasenose, Pembroke, and Keble colleges, and I am grateful, in particular, to Dave Leal, Adam Humphreys, Stephen Whitefield, and Ali Rogers for the invaluable support they have offered in that time. In addition, I owe my thanks to the library staff at Nuffield, Keble, and Brasenose colleges, where much of the research presented here was conducted.

Inevitably, there are a few people to whom I owe intellectual and moral debts which are hard to quantify, and it isn't always easy to see where one kind of debt ends and the other begins. In no particular order, I thank Michael Laver, Eddie Hyland, Jac Hayden, Rob Jubb, John Filling, David Birks, Sam Kiss, Tony Taylor, Phil Smith, Gráinne de Bhulbh, my parents, and my family. They have many virtues, but it is perhaps their patience upon which I have relied most.

IJC, 05 July 2022

Contents

Introduction 5

Part I – Why Not Contractarianism?

Chapter 1 **25**

§1.1 Liberal political institutions, normativity, and political contractarianism 26

§1.2 Hypothetical contracts and prescriptive force 27

§1.3 Moral and prudential reasons 30

§1.4 Categorising social contract theories 33

§1.5 Contractarianism and moral realism 35

§1.6 Contemporary contractarianisms 39

§1.6.1 Gauthier's agnostic contractarianism 42

§1.6.2 Moehler's two-level contractarianism 48

§1.6.3 Vanderschraaf's conventionalist contractarianism 57

§1.6.4 Muldoon's experimental contractarianism 67

§1.7 A contractarian conception of justice 76

Chapter 2 **79**

§2.1 The scope of politics 80

§2.2 Two levels of political reasoning 85

§2.3 Reasons and motivation 88

§2.3.1 Effective reasons 89

§2.3.2 Pervasive and persistent disagreement 92

§2.3.3 Convergence and consensus 99

§2.4 Moral and prudential reasons revisited 103

§2.4.1 Self-directed reasons 104

§2.4.2 Priority and redundancy 107

§2.4.3 Outcome-oriented and action-oriented reasons 112

Chapter 3 **121**

§3.1 Two conceptions of the problem of moral disagreement 121

§3.2 Contractualism *contra* contractarianism 129

§3.2.1 Constituency of justification objections 129

§3.2.2 Political-in-the-wrong-way objections 136

Part II – Justice as Mutual Advantage

Interface 145

Chapter 4 **168**

§4.1 Two conceptions of liberalism 169

§4.2 The authority-based conception and contractarianism 181

Chapter 5 **188**

§5.1 Preferences and rationality 188

§5.2 Formal constraints on preference sets 191

§5.3 Objections to completeness 194

§5.4 Objections to transitivity 205

§5.4.1 Discernment objections 206

§5.4.2 Multidimensionality objections 209

§5.4.3 Money-pump redux 214

§5.5 The circumstances of contractarian justice 215

§5.5.1 Five facts about the world 216

§5.5.2 Evaluating the facts 219

§5.6 Where we are 226

Chapter 6 **228**

§6.1 The bargaining problem 228

§6.2 The egalitarian bargaining solution 233

§6.3 The Nash and Kalai-Smorodinsky solutions 236

§6.4 A procedural comparison between Nash and Kalai-Smorodinsky 241

§6.5 An axiomatic comparison between Nash and Kalai-Smorodinsky 247

§6.6 A bargaining solution for the contractarian model 257

Chapter 7 **260**

§7.1 Rationality, individualism, and universalism 263

§7.2 Freedom, equality, and austerity 264

§7.3 Equal and ideal rationality 267

§7.4 Equal and ideal information 268

§7.5 Equal endowments and the contractarian baseline 272

§7.6 Equal prospects 285

§7.7 Neutrality 293

§7.8 The scope of contractarian institutions 295

Chapter 8 **302**

§8.1 The coalition problem 303

§8.2 Unsuccessful responses to the coalition problem 305

§8.2.1 The pragmatic response 305

§8.2.2 The natural core response 307

§8.2.3 The cooperative infrastructures response 311

§8.3 The side payments response 318

§8.3.1 Bargaining power as a function of strategic power 318

§8.3.2 Justifying side payments 321

§8.3.3 Side payments in asymmetric contexts 325

§8.4 Global rationality 331

§8.5 A basis of (strategic) equality 339

Conclusion 345

Bibliography 347

Figures

Figure 1 A geometric description of the bargaining problem 232

Figure 2 A geometric description of the Nash bargaining solution 237

Figure 3 A geometric description of the Kalai-Smorodinsky bargaining solution 239

Games

Game 1 A 3-player symmetric divide-the-dollar game 303

Game 2 A 3-player asymmetric game with an essential player (the Macquarrie-O'Rourke-O'Reilly game) 308

Game 3 A 3-player symmetric divide-the-dollar game with variable payoffs 313

Game 4 A 3-player asymmetric game with variable payoffs 314

Game 5 A 4-player asymmetric game with a dummy player 321

Game 6 A 4-player asymmetric game 325

Tables

Table 1 Payoffs for Game 4, given initial choice of coalition 315

Table 2 Payoffs for Game 4, including bargaining floors 327

Introduction

The problem

My world is imperfect. Looking around, things are not arranged as I would like them to be. Sometimes, it is within my power to alter existing arrangements. If I am thirsty, and wish to alter the world so that I am not, I can go to my kitchen, pour clean water from a tap, and drink. In doing so, I move the world ever so slightly closer to my idea of perfect.

My world is unjust. There are many people who do not have the access to clean water that I have. Altering the world to correct this imperfection is more difficult than pouring water from a tap. By myself, I lack the power to make the world just. I can only realise justice by coordinating my actions with many other people.

My world is divided. Most people disagree with my conception of justice. My best efforts to persuade them of the truth of my conception fail. I lack the power to simply force everyone to act justly. This presents a problem: How can I best pursue my conception of justice when the people upon whom its realisation depends cannot be persuaded that it is correct?

This is the question I set out to answer. Note, first, that I am not attempting to say what justice demands in an ideal world. I am taking for granted that the world is non-ideal in at least one respect: The people in it hold conflicting conceptions of justice, and no one has the power to impose the correct conception of justice on everyone else. This is not the same as supposing that disagreement about justice is ineradicable, or that trying to persuade each other that one conception of justice is better than another is a waste of time, or that there is no objective truth about justice. I will propose a theory for people who live in a world where there is disagreement. Doing so is consistent with believing that we need not always live in such a world.

The assumption that there is disagreement in the world is not unusual in contemporary political philosophy. The challenge posed by reasonable disagreement, for example, motivates much of John Rawls's later work.[1] Philosophers following Rawls have tried to work out what the correct conception of justice is, paying due attention to how the existence of (reasonable) disagreement affects the content of this conception. My concern is different. Suppose the Rawlsian project succeeds, and we identify the correct conception of justice. There is a further problem. In order to implement a conception of justice, we must either be able to persuade many people that the conception is correct, or have sufficient power to impose it in the face of widespread resistance. Given the degree of disagreement in the world, and our limited powers, we will often find that neither of these conditions is met. What distinguishes my project, then, is a concern with the conjunction of two things: the existence of disagreement, and our practical dependence on the support of the people with whom we disagree.

One might think that a problem of this sort is not the concern of political philosophers. Philosophers are not rhetoricians or tacticians. Their job is to correctly identify what justice demands. Once this has been done, the further problem of how one goes about realising justice is a task for people with other kinds of expertise.

There is plenty of truth in this thought. Many problems which arise when implementing a conception of justice will be particular to a context and will not be especially amenable to resolution by philosophical analysis. Nonetheless, I will contend that there is a (more or less) general theory of coordination in conditions of disagreement to be had, and that this theory is the concern of political philosophy. Philosophically, the problem is this: Suppose that people act on the basis of reasons.[2] What reasons can I give, to a person who disagrees with my conception of justice and who holds a conflicting conception of her own, which will motivate someone who acts on the basis of reasons to help me realise my conception of justice? While

[1] 2005. Compare Barry, 1995; Gray, 2000; Gaus, 2011; Larmore, 2013; Moehler, 2018: 1-2, 33.
[2] Korsgaard, 1996: 92-102, 112-113.

people do not always act on the basis of reasons, and the reasons which motivate particular people in particular circumstances will often be context-dependent, identifying a set or class of reasons with which this question can be answered is a recognisably philosophical problem.[3]

Disagreement about justice is an instance of the more general phenomenon of moral disagreement. Sometimes, when we disagree about moral principles, we can resolve our disagreement by appealing to higher values. Thus, we might settle a dispute about the permissibility of abortion by appealing to democracy – by voting. Or, we could resolve a conflict between individual freedom and democratic authority by appealing to the value of utility. We might settle a religious disagreement by appealing to religious toleration – by agreeing to differ. One response, then, to disagreement about justice is to identify some higher-order value, principle, or procedure which tells us how to respond in cases of first-order disagreement.

Disagreement about justice presents a special problem, however, because of the kind of thing that justice is. I understand justice to be a moral principle that takes priority over other moral principles.[4] I intend this to be a conceptual claim rather than a substantive one. If you were to argue that principles of beneficence take priority over principles of justice, I would not say that you have made a substantive mistake (only). The problem is not that you have overestimated the importance of beneficence. Rather, on my view, you have made a category mistake. For you, I would say, justice *is* beneficence, because beneficence is the moral principle which you take to have priority over other principles.

On this understanding, disagreement about justice presents a special problem because justice, like truth, is uncompromising.[5] Disagreements about justice cannot be resolved by appealing to higher moral principles because justice is supposed to

[3] Margalit, 2009: 5-6, 12.
[4] Rawls, 1999a: 3-4; Rawls, 2005: 137-138. Compare Steiner, 2013: 232-235. See §1.7, *infra*. For another view, contrasting the priority of justice with that of peace or order, see Moehler (2018: 211-213).
[5] Rawls, 1999a: 4. Compare Margalit, 2009: 1-8, 12, 113-115.

be the highest of our moral principles. Justice defines the *limits* of democracy, utility, and toleration. It determines those things which democratic majorities must not take from us, which must not be sacrificed for utility, and which we must not tolerate.

Four possible solutions

I am concerned with disagreement of the highest order.[6] Thus, resolving disagreement by appealing to higher-order moral principles is impossible. This leaves us with four options. First, we could look for a resolution *within* our theories of justice. If a principle for resolving disagreement is itself a principle of justice, then it can have sufficient priority to overturn other principles of justice. Second, since the problem is driven specifically by moral disagreement, we could try to resolve our disagreement by appealing to a non-moral principle. Third, we might simply have to accept that it is not possible to resolve our disagreements about justice. Fourth, we might be inclined to reject one of the premises with which I began. It strikes me as implausible to deny that there is profound disagreement about justice, but one might doubt the conjunction which I have assumed. That is, one might think there is sufficient agreement about *some* principles of justice, among a sufficiently large proportion of people, that those people will have the power to implement the shared components of their conceptions of justice without having to rely on the support of those with whom they disagree. In this way, one might argue for implementing a restricted conception of justice. In order to attract sufficient support, and hence secure sufficient power to enforce a conception of justice against all dissidents, the area of agreement would be narrow, perhaps only including the most fundamental principles (prohibitions on most killing, maiming, and torture, say).

Each of these options is problematic in some way. Consider the possibility of simply giving up on resolving disagreement. By hypothesis, we lack the power to enforce justice by ourselves, and the capacity to convince others that our conception of

[6] Compare Moehler, 2018: ix, 33. Moehler identifies disagreements of this kind as "cases of conflict" (*ibid*.: 15-16; see §1.6.2, *infra*).

justice is correct. Thus, this option would leave us unable to realise our conceptions of justice. More precisely, it would mean we could not fully realise our conceptions of justice, and that we could partially realise a conception of justice only to the extent that our individual powers allow. In most cases, this will be to a very limited extent. Worse, if justice is something we take to be coercively enforceable, and if the enforcement of justice is among our highest-order moral duties, our inability to resolve disagreement looks like a moral catastrophe. We will find ourselves in Hobbes's war, as each of us attempts to violently impose our conception of justice to the best of our abilities.[7] This is not an attractive conclusion.

Alternatively, consider the possibility of securing a consensus on a restricted conception of justice. Suppose there is a society in which some of the people are Rawlsian liberals, most of the rest are Nozickian libertarians, and a remaining handful are unrepentant Nazis. The liberals and libertarians can agree on certain rights, and between them have enough power to suppress any attempts by the Nazis to violate these rights. Nonetheless, there is profound disagreement between the two larger groups on purported rights to equality of opportunity and property.

If neither group has the power to fully enforce its conception of justice, it is not obvious that the solution can be a regime which enforces what the two groups' conceptions of justice have in common. Suppose Jack, being less advantaged than Bob, seizes some resources Bob considers to be rightfully his. Bob agrees that Jack has a right against being coerced in the normal run of things, but Bob does not think that Jack has a right against being coerced when he is acting unjustly. Are we to understand liberals and libertarians to overlap on the belief that Jack has a right against being coerced, or that he has a right against being coerced unjustly? Clearly, the latter will not do, since there is disagreement about the full extension of justice. But the former will not do either, since neither group thinks that Jack has an unqualified right against being coerced.

[7] *Leviathan*, Chapter XIII (1996: 86-90). Compare Gaus, citing Kant and Hobbes (2003: 16-17, 63-64); Barry (1995: 30-31, 39); Abizadeh (2011: 306, 308-309); Moehler (2018: 104).

This is just an example, but it illustrates the difficulties with the restricted consensus option. Moreover, in the scheme of ideological disagreement, liberals and libertarians have a lot in common. If such difficulties arise in this kind of case, they seem likely to be worse in cases involving a wider range of perspectives. And, the narrower the range of groups which can be included in a consensus, the less likely the groups in the consensus will have sufficient power to implement a restricted conception of justice without the assistance of groups outside the consensus.

At this point, it might be tempting to say that the liberals and libertarians could coincide on a method of resolving differences peacefully – on a democratic procedure, for instance, duly constrained by a bill of rights. But, as I have already argued, this cannot be the case if recourse to the procedure is not itself endorsed by a principle of justice. Given the priority of justice, the output of any procedure not so endorsed would lack the standing to arbitrate a dispute over competing claims of justice. If we take this approach, our supposed fourth option is simply a version of our first.

Suppose, then, we look for a resolution of the problem of disagreement about justice within our theories of justice themselves. We will need to be careful how we specify the claim we are making. For example, it will not be sufficient to say that, when there is disagreement about justice, our theory of justice tells us to resolve the disagreement democratically. The problem is precisely that others might disagree that disagreement ought to be resolved democratically. This is true of all principles of justice which are used to resolve disagreement. Unless there happens to be a broad consensus on a principle of justice for resolving disagreement about justice – and it isn't clear to me why disagreement would be conveniently absent in this part of a theory of justice but not in others – appealing to such a principle will fail to solve the problem of disagreement.[8] Because it is a principle of justice, the principle will have the standing to challenge competing principles of justice. But because it is itself subject to disagreement, the principle will not motivate the people whose

[8] I take up this point in more detail in §2.3.

support we need to implement it.

Finally, suppose we try to resolve our disagreement about justice by appealing to some non-moral principle.[9] One might immediately think this option is not viable, for the same reason that we could not appeal to a higher-order moral principle to resolve disagreement about justice. If realising justice is our highest moral end, we might think it will take priority over any non-moral principle. But this doesn't quite follow. That justice trumps all other moral considerations does not imply that what we ought to do, all things considered, is pursue justice, unless one also thinks that our moral reasons always trump our non-moral reasons, that all of our reasons are moral reasons, or that justice trumps our non-moral reasons as well as our other moral reasons (even if not all of our moral reasons trump all of our non-moral reasons). Still, many people accept one of these additional claims. Thus, any non-moral solution to the problem of disagreement will be controversial. To that extent, it will fail to motivate the support of the people we need to implement it.[10]

Political contractarianism and *modus vivendi*

The solution I propose to the problem of disagreement about justice relies on combining aspects of these latter two options. It is an example of what is called a *modus vivendi* theory.[11] Suppose that people care, not just about fully implementing their conceptions of justice, but about advancing them to the greatest extent possible. Suppose further that it is possible to find a compromise between competing conceptions of justice which, while not realising any conception fully, realises each to a greater extent than any would be realised in the absence of the compromise. I will then say that I have a moral reason, based on justice, to maximally advance my

[9] The main example I have in mind is a principle grounded in prudential reason, as distinct from moral reason. There may be other examples. For further discussion see §1.3 and §2.4.1.
[10] See §2.4.
[11] See §3.2.2. For helpful discussions of *modus vivendi*, see Gaus (2003: 58-59), Rawls (2005: 39-40, 147), Quong (2011: 17, 164), Sleat (2013: 90-91, 94-109), Moehler (2018: 176-179), and Vallier (2019: 107-110). Sleat reviews work by Neal (1997), Gray (2000), Margalit (2009), Horton (2010), McCabe (2010), and Rossi (2010).

conception of justice, and that compromising with others is the most effective way to do so. The reason I give, to the people with whom I propose to compromise, to motivate them to help me promote my conception of justice, will not be that they ought to help me promote my conception because it is correct. Rather, the reason will be that they ought to help me promote my conception of justice because, if they do, I will help them to promote their conceptions of justice.[12]

I will describe this solution as *political contractarianism*. A 'contractarian' theory tells me that what I ought to do is whatever advances my ends to the greatest extent possible. When I need your coordination to advance my ends, the theory tells me to act in such a way as to advance your ends if doing so is an effective means of motivating you to act so as to advance mine. The central idea is that of a contract – an agreement between two or more parties, who do not take an intrinsic interest in one another's ends, to nonetheless coordinate their actions in such a way as to advance the ends of each party.

My account is 'political' in contrast to moral contractarianism.[13] Moral contractarianism takes morality to be constituted by the compromise which emerges from rational agents pursuing their ends. Morality, on this view, is *derivative* of prudential reason – morality is enlightened self-interest. In contrast, political contractarianism sees morality as *autonomous* of prudential reason. Morality, on the political view, is among the most important sources of an agent's ends, which she then pursues in an instrumentally rational way. Political contractarianism thus models citizens' conceptions of justice as ends which they hold, and then prescribes a rational compromise which allows each citizen to advance her conception to some degree greater than she would be able to advance it in the absence of the compromise.[14] The theory draws on principles from within our conceptions of justice, because it supposes that the reason we have to compromise is that justice itself requires us to advance justice to the greatest extent we can. However, in

[12] Compare Korsgaard, 1996: 132-134; Vanderschraaf, 2019: 274-276.
[13] See Gauthier, 1986; also §1.4, §1.5, *infra*.
[14] Barry, 1995: 30, 32; Moehler, 2018: ix.

constructing its model of rational compromise, the theory also draws on a non-moral principle rooted in prudential reason.[15]

By proceeding in this way, I hope to avoid the difficulties I identified earlier. The problem with appealing to a non-moral principle to resolve disagreement was that our duty to advance justice could have priority over any such principle. If our reason to employ the non-moral principle as a method of resolving disagreement is itself grounded in our commitment to justice, we can sidestep this priority problem. On the other hand, the problem with appealing to a principle of justice to resolve disagreement about justice was that there might be as much disagreement about the principle for resolving disagreement as about any other principle of justice. I hope to evade this reflexivity problem by picking a principle for resolving disagreement about which there can be no sensible disagreement: instrumental rationality.[16] My claim is that, in a world characterised by disagreement about justice, we ought to coordinate with one another to implement a conception of justice which represents a rational compromise between all of the pure conceptions of justice held by those parties whose agreement is necessary to implement the compromise conception. We will not *share* reasons for coordinating to implement the compromise conception of justice, but our individual reasons will *converge* on implementing the compromise.[17]

For such an argument to be successful, two key difficulties have to be overcome. First, it may seem that there is not much more to say than I have already said. I claim that, in the non-ideal circumstances I have described, we ought to do whatever we can to advance our conception of justice, compromising where necessary. Just on its face, this claim looks trivial. However, what has distinguished contractarianism from other *modus vivendi* accounts has been a willingness to use formal methods to give more determinate content to the compromises it prescribes.[18] While no *modus vivendi* theorist would deny that the content of a compromise must be to some extent

[15] Compare *ibid*.: 19.
[16] *Ibid*.: 100, 105, 121-122.
[17] See §2.3.3.
[18] See, for example, Gauthier (1986), Binmore (1994; 1998), or Vanderschraaf (2019).

context-dependent, I will argue that a contractarian theory, properly conceived, is about as determinate in its prescriptions as Rawls's principles of justice. There is plenty of room for contextual variation, but this variation is constrained by a set of contractarian principles.[19]

The second problem with any *modus vivendi* theory is that, from the point of view of a partisan of any particular conception of justice, acquiescing in a *modus vivendi* is a compromise with injustice. Earlier, I imagined a society composed mostly of liberals and libertarians. We can suppose that a society built around a compromise between those groups would not be catastrophically unjust, in the substantive sense that it would respect basic rights. Nonetheless, the distribution of conceptions of justice which we find in the people around us is a matter of contingent fact, and it is this distribution which determines the compromise a *modus vivendi* theory prescribes. If we find ourselves a small minority of liberals in a society mostly made up of Nazis and religious fundamentalists, the substantive political institutions recommended by a contractarian theory could be grossly, unacceptably unjust.[20]

This is, I take it, the central challenge to political contractarianism. Some *modus vivendi* theorists attempt to respond to such worries by building basic moral constraints into their models.[21] Unfortunately, doing so renders their arguments unstable, in two ways. First, once one has conceded that there can (or ought to) be agreement on some basic moral commitments, the question is why there cannot (or ought not to) be more extensive agreement. Any response looks like it will arbitrarily divide those moral propositions which I think are obviously basic from those which

[19] Gauthier, 2013: 620. For further discussion, see §1.6.1, *infra*.
[20] Compare Vanderschraaf, 2019: 304-305, 310.
[21] See Sleat, 2013: 99-100; Gray, 2000: 6, 8, 107-108; Gaus, 2003: 65-67; Margalit, 2009: 1-7, 89-119; Baderin, 2014: 139-140; Moehler, 2018: 5, 176-179. Moehler (*op. cit.*: 3, 148-155, 176-177) attributes an approach of this kind to Gaus (2011), Vallier (2014), and Muldoon (2016), as well as to Gray (2000) and McCabe (2010). Vanderschraaf (2019: 309-311, 320, 322) describes the same strategy in theorists as diverse as Rawls, Harsanyi, Scanlon, Binmore, and Gauthier (some of whose constraints, of course, are rather more than basic). Some theorists – following Williams (2005; see also Horton, 2010: 438-439; Baderin, *op. cit.*) – may be inclined to make this a conceptual claim about the nature of political society rather than a substantively moral commitment. I doubt this strategy can avoid the problem entirely, but I will discuss a similar move in §2.1, *infra*. Compare Kiss, 2016: §4.5.

you think are similarly basic. Second, flinching from the implications of a *modus vivendi* theory in an extraordinarily vicious world, but not in an ordinarily vicious world, makes no sense. It would be terrible to find ourselves caught between Nazis and religious fundamentalists. But, if that is where we find ourselves, it is no less true that we will have to decide what to do.[22] From the perspective of a *modus vivendi* theory, wishing that more people agreed with us on at least basic moral commitments is no more or less helpful a response in such circumstances than wishing for complete agreement on Rawlsian justice is in less desperate straits.

So, a concessionary response is unpersuasive – granting basic moral constraints undermines the core commitments of a *modus vivendi* theory.[23] The central challenge remains: In certain circumstances, political contractarianism could tell us we ought to do profoundly unjust things. If the theory is inconsistent with our most important convictions about justice, why shouldn't we reject it?

My reply to this challenge has two stages. First, accepting political contractarianism does not commit one to rejecting any *ideal* theory of justice whatsoever, because the primary concern of political contractarianism is a world which is non-ideal in at least one important respect. We shouldn't be too quick to insist that we can read off what justice demands in truly terrible circumstances from our ideal theories.[24] Once we properly consider those circumstances, we may find that justice permits – even requires – us to choose evils, so long as they really are lesser.[25]

Still, perhaps the challenge can be sharpened. Perhaps the real worry is that political contractarianism will prescribe profoundly unjust actions, not only in truly terrible circumstances, but in perfectly ordinary ones too, and specifically in the circumstances in which we tend to find ourselves as citizens of modern, liberal democracies. Given that contractarianism admits no moral constraints, and is based on a balance of power which we have no reason to believe will be egalitarian, how

[22] Compare Margalit, 2009: 13-14, 113-115, 134-140, 175-197.
[23] Rossi, 2010.
[24] Enoch, 2015: 125.
[25] Margalit, 2009: 113-115, 134-140.

can it endorse core liberal commitments such as equality under law, equal basic rights, and democracy? The second stage of my reply argues that, in a wide variety of circumstances – wider than has previously been appreciated – contractarianism prescribes the implementation of a liberal egalitarian conception of justice, roughly in line with Rawls's principles of justice. Thus, in ordinary circumstances, if we are liberals, we are not forced to choose between contractarianism and our substantive conception of justice.

It may seem that there is something alchemical in this claim. Much contemporary liberal theory takes it as axiomatic that one cannot get liberal justice from such leaden material as prudential reason. Certainly, I do not claim that doing so is straightforward. But the attempt is not as obviously hopeless as it is usually taken to be. First, note that I am not trying to derive an 'ought' from an 'is', or a moral ought from a prudential ought. For *political* contractarianism, my claim is that the prudential reasons most citizens have will often coincide with the moral reasons liberals take themselves to have, whether or not most citizens are themselves liberals. If this claim can be defended, it will answer one of the most persistent challenges to Rawlsian forms of liberalism: It will justify liberal political institutions to people who do not already accept the moral premises upon which Rawls's theory relies.[26] Second, this convergence of citizens' prudential reasons on liberalism is not simply a coincidence. The idea of a prudentially-grounded *modus vivendi* has been a thread running through liberal theory from its origins.[27] As an aspect of liberal thought, it has drifted in and out of favour, but has never disappeared entirely. It has often been recognised that, if one assumes roughly equal power among citizens, something like liberal political institutions follow as a *modus vivendi*.[28] The key problem has been justifying the egalitarian premise. Asserting it as a moral minimum is problematic, as I have argued above; and it seems obviously false as an empirical claim about the world. In the absence, then, of any other reason to accept the egalitarian premise, contemporary liberals have thought *modus vivendi* a dead end.

[26] Quong, 2011: 138-140, 158-160.
[27] Gray, 2000.
[28] Quong, 2011: 17.

Thus, one of the most important steps in my argument will be defending a contractarian basis of equality.[29] While I will have much more to say on the topic, I can give an outline of the idea now. Consider a world in which there are three people – *A*, *B*, and *C*. Suppose that these people are all interested in winning for themselves as much of some prize, *P*, as possible. Suppose that gaining *P* requires contributing 10 units of effort to a task. *A* is the strongest person; she can contribute 9 units of effort. *B* can contribute 7 units. *C* is weakest and can contribute just 3 units. No one can gain *P* alone, but any two people working in combination can gain *P*, and then divide it as they please.

The important aspect of this scenario to grasp is that any two-person agreement is vulnerable to being disrupted by the third, excluded person. If *A* and *B* come to some agreement to divide *P*, *C* can always offer one of them an alternative deal which represents an improvement on their original agreement. *Modus vivendi* theories in general, and contractarian theories in particular, rely on a balance of power. They are often thought to be inegalitarian because people differ in their abilities, as *A*, *B* and *C* do. But, for the purpose of a contractarian theory, *power* is not equivalent to strength, or productive capacity, or number of supporters. When parties lack the ability by themselves to achieve an end such as implementing a conception of justice, power is a function of one's ability to combine with others and of one's ability to prevent others from combining. Although the parties in a case like the *A*-*B*-*C* scenario are not equal in their productive capacities, each of them equally possesses the ability to make an offer which disrupts the formation of a coalition which excludes them. I will argue that this ability is decisive in determining the allocation of the burdens and benefits of coordination. It is not a given, of course, that in any particular case everyone will possess the ability to make a disruptive offer. But, crucially, one either possesses the ability or one does not; one does not possess it in degrees. It is thus what Rawls calls a *range* property.[30] As such, it can serve as a basis of equality.

[29] Rawls, 1999a: 441-449. Compare Vanderschraaf, 2019: 295.
[30] Rawls, 1999a: 444-445.

Two things I will not discuss

Before I outline the structure of the argument to come, let me note a couple of things I will not discuss. Several of the themes I pursue – disagreement, order, action-guidance, motivation – play a prominent role in a set of recent writings which have been gathered under the label of 'political realism'.[31] While I occasionally draw on this literature, I do not engage with it directly, for two reasons. First, my main focus is a contractarian response to these themes, and contractarianism has been conspicuous by its absence in realist writing. Second, some realists' objections to contemporary liberal theory have targeted what they take to be its excessive abstraction and idealisation. I don't share this methodological concern. The problem I address relies on the world being non-ideal in one respect. But, in investigating the problem, I will help myself to a range of idealisations and abstractions. I find these are useful tools, without wishing to claim that they are the only tools available, or that there are no costs associated with using them.[32]

A contractarian theory typically rests on three pillars: an account of rational agreement on a pattern of coordination, an account of rational compliance with such agreements, and a conception of the starting point from which agreement proceeds.[33] My second major omission is that, while I will say quite a lot about the first and third of these, I will hardly mention the second.[34] This is despite the fact that the compliance problem is perhaps the most prominent controversy in contemporary contractarian theory.[35]

---

[31] These writers trace their tradition through Thucydides, Machiavelli, Hobbes, and others (Sleat, 2013: 2). The contemporary touchstones are Bernard Williams (2005) and Raymond Geuss (2008). Sleat (2013: 17) provides a useful bibliography, as does Kiss (2016).

[32] For a study of realism as a method in political philosophy, see Kiss (*ibid*.). For reflection on contractarianism as a method, see the Interface, *infra*.

[33] Gauthier, 1986: iv.

[34] Chapters 6 and 8 are concerned with rational agreement; §7.5 is concerned with the starting point, but touches briefly on compliance.

[35] Gauthier, 1986: 157-189; Morris, 2013: 599-600. What is probably the most influential statement of the problem is given by Hobbes, in *Leviathan*, Chapter 15 (1996: 101-102); but compare Plato, in *Republic*, 358b-367e (1993: 45-56); and Hume, in *A Treatise of Human Nature*, §3.1.5 (1984: 573). See also Kavka (1983), the essays in Vallentyne (1991a: 177-322), Binmore (1994: 26-28, 173-182), Gauthier (1994; 2013), Barry (1995: 33-39), Coram (2001: 42-43), Bratman (2013), Finkelstein (2013), MacIntosh (2013), Thoma (2015: 3338, 3349-3350), Moehler (2018: 46-49, 51, 56-58, 103, 107, 142-145, 172-174), Vanderschraaf (2019: 226-228ff). Vanderschraaf provides a particularly

I set aside compliance for three reasons. First, and most straightforwardly, due to considerations of space. Second, at the centre of the compliance debate is a controversy between David Gauthier and his critics. I am, more or less, persuaded by the critics. I have little to add to their objections. Contractarianism does need a theory of compliance, but in this project my attention is elsewhere. Finally, the compliance problem takes on slightly different forms depending on whether one is working with moral or political contractarianism, and it is a more pressing concern for moral contractarians.[36] Political contractarians have resources to address the problem which moral contractarians lack. For instance, in political contractarianism, agents can have moral reasons to comply with agreements which the amoral agents of moral contractarianism lack. So, the fact that I reject Gauthier's account of compliance does not require me to reject contractarianism outright.

Outline of an argument

How can I best pursue my conception of justice when the people upon whom its realisation depends cannot be persuaded that it is correct? One can read this question in either a general or a specific way. Generally, the question is asking what *kind* of theory gives the best response to the problem of coordination in conditions of disagreement. I argue that political contractarianism is our best answer. However, we can also read the question as asking what the specific *content* of a contractarian compromise would be – what specific political arrangements (if any) such a theory would endorse. I argue that, in the circumstances of contemporary societies, contractarianism endorses *liberal* political institutions.

My discussion is thus divided into two parts. Part I defends political contractarianism as a general response to the problem of coordination in conditions of disagreement.

---

illuminating game-theoretic discussion of the problem (*op. cit.*: 200-215, 233-236, 240, 242-259, 262-267).

[36] Barry, 1995: 33-36; Moehler, 2018: 172-174.

Chapter 1 defines political contractarianism. The key point here is that a *political* contractarianism is motivated by moral disagreement, not by moral scepticism. It is thus compatible with (though it does not require) moral realism. Chapter 2 presents my main argument for political contractarianism – what I will call the argument from effectiveness. The idea will be that solving the problem of moral disagreement requires giving others motivationally effective reasons for coordination, and that prudential reasons are better candidates for this role than moral reasons. Even if we agree that we have moral duties which conflict with the prescriptions of a contractarian theory, these duties do not take priority over the contractarian prescriptions in our practical reason, because the contractarian prescriptions are ultimately underpinned by our commitment to promoting justice. In Chapter 3, I defend my contractarian account against five objections drawn from a Rawlsian, contractualist perspective.

In an important sense, the conclusion of Part I is purely formal. It says that we have reason to accept whatever political institutions are recommended by a contractarian theory, *regardless* of their substantive content or our moral commitments, because the theory recommends the best (by our lights) that we can get, given the constraint that we need the coordination of others but cannot convince them of the truth of our conception of justice. This formal conclusion leaves open the possibility that a contractarian theory won't have any usefully specific content about the sort of political institutions it endorses; and further, that it could recommend (what we would consider to be) profoundly unjust political institutions, given certain contexts. However, in Part II I argue that, in the circumstances of contemporary societies, contractarianism will specifically recommend liberal (and, because liberal, just) political institutions.

I begin, in Chapter 4, by arguing for a particular, *constructivist* conception of liberalism. That is, I argue that liberal political institutions are properly *defined* as those institutions which would be chosen by suitably idealised agents in a certain hypothetical choice situation. In the rest of Part II, I build a contractarian model which, I argue, (i) meets the conditions I have set out for constructivist liberalism

and (ii) includes no morally-motivated constraints on the agents or their choice situation. Chapter 5 defends my conception of the agents and their choice situation. Chapter 6 discusses bargaining theory. I identify five characteristics of agents and their circumstances which, taken together, determine the distribution of the burdens and benefits of social coordination. The contractarian model prescribes liberal political institutions when its agents are equal in their rationality, information, endowment, prospects, and bargaining power. In Chapter 7, I argue that we should understand agents in the contractarian model to be approximately equal in the first four of these ways. In Chapter 8, I argue that we should also understand them to be equal in bargaining power, with a few important qualifications. I conclude that, properly understood, in the circumstances of contemporary societies, political contractarianism endorses liberal political institutions.

Part I

Why Not Contractarianism?

Chapter 1

Political contractarianism is a kind of social contract theory. The basic idea behind a social contract theory is that we can work out how we ought to arrange some aspect of our social world by asking how we would hypothetically agree to arrange it if (say) we were thinking clearly, were well-informed, and were bound by our agreements.

Social contract theories differ according to how they refine this idea, but two dimensions of refinement are particularly salient. First, the *basis* of agreement can vary. Social contract theories assume that people act on the basis of reasons. There are different kinds of reasons, and using one set rather than another as the basis of agreement generates a range of different theories. Second, the *object* of agreement can vary. Social contract theories can be applied to various aspects of our world. In political contractarianism, the basis of agreement is prudential reasons, and the object of agreement is a set of normative principles to regulate political interaction.

In this chapter, I clarify some basic concepts, and answer some preliminary objections. In §1.1, I give an initial characterisation of my task. In §1.2, I defend my use of hypothetical agreement. In §1.3, I distinguish prudential reasons from moral reasons.

In §1.4, I distinguish political contractarianism from three other types of social contract theory: moral contractarianism, moral contractualism, and political contractualism. Each of these differs from political contractarianism either on the basis of agreement (moral versus prudential reasons), on the object of agreement (moral versus political principles), or on both. In §1.5, I argue that we may wish to reject moral contractarianism, but that rejecting it does not imply we should also reject political contractarianism. In §1.6, I examine a number of contemporary

contractarian accounts in order to situate my account in the existing literature and to distinguish it from that literature. Finally, in §1.7, I discuss the relationship between 'a set of normative principles to regulate political interaction' and 'justice'.

§1.1 Liberal political institutions, normativity, and political contractarianism

My ultimate aim is to defend a contractarian account of liberal political institutions. I describe this account as political contractarianism. When I say 'political institutions', I have in mind the state and its component parts.[1] By 'liberal' political institutions I mean a set of political institutions that are in some sense shaped by the values of freedom, equality, and neutrality.[2] Note also that my project is *normative*. I argue that we *ought* to create and sustain liberal political institutions.[3]

My account of liberal political institutions is 'contractarian'. This means that my argument takes a particular form. First, it is a social contract theory. I argue that certain normative principles regulating political interaction are justified because rational agents, in the right hypothetical circumstances, would agree to be governed by those principles.[4] Second, it is a particular kind of social contract theory: a contractarian theory. This means that what justifies each agent in agreeing to the principles is the fact that she has *greatest prudential reason* to choose those principles, as opposed to any others. Third, the object of the agreement is a set of principles to regulate political interaction. My claim will be that we ought to create and sustain liberal political institutions because those are the only institutions consistent with the principles which would be chosen, on the basis of prudential reasons, by rational agents seeking to regulate their political interaction with one another.

[1] For my conception of what counts as political, see §2.1. I leave the content of what an institution is fairly open (Rawls, 1999a: 47-50).
[2] For further detail, see §4.1.
[3] Compare Gauthier, 1986: 2, 3.
[4] Compare Moehler, 2018: 11; Vandershraaf, 2019: 198-200, 209. For my specification of rational agents and their circumstances, see Chapter 5.

§1.2 Hypothetical contracts and prescriptive force

In discussing social contracts, it is typical to distinguish two conceptions of agreement. Sometimes, agreement is understood as an actual historical event. Alternatively, we can conceive of a social contract as a hypothetical agreement – an agreement into which we say an agent *would* enter, given certain conditions. Contemporary theorists tend to favour hypothetical agreement.[5] For my purpose, what it is important to get clear is the *prescriptive force* of hypothetical agreement. Why should we think that the fact that a particular agent would agree to *X*, in some way justifies *X*?[6]

There are two reasons one might have for doubting that hypothetical choice – and hence hypothetical agreement – has prescriptive force. First, one might ask why the fact that I would choose to do some thing has any bearing whatsoever on whether or not I ought to do that thing. That I would lend you my car if you asked to borrow it does not seem to justify me lending the car to you.

This worry is based on a misunderstanding of the role hypothetical choice plays in a social contract theory. Hypothetical choice theories are designed to tell real people what they *should* do by asking what those people *would* do *if they were idealised* in certain respects (if they were thinking clearly, were well-informed, etc.). Different theories will idealise in different ways, but the basic motivation for doing so is the same – to make the agent in question responsive to whatever reasons the theory takes to be important. Thus, the fact that I would lend the car to you, if I were fully informed about the circumstances, justifies me lending the car to you when asked in the real world, where I am not fully informed.

---

[5] Rawls, 1999a: 11-12; Gauthier, 1991: 25; Gauthier, 1997: 132-133; Gauthier, 2013: 616-619; Thoma, 2015: 3338; Thrasher, 2019: 429-430; Vanderschraaf, 2019: 209. Moehler (2018: 11, 33, 178-179, 217), by contrast, endorses actual agreement. For standard objections and defences, see Dworkin (1989), Waldron (1987: 138-146), Stark (2000), Larmore (2003: 369-371), Rawls (2005: 24-25), D'Agostino, Gaus, and Thrasher (2017: §3.1), or Enoch (2015: 127-130).

[6] See Korsgaard's discussion of 'the normative question' (1996: 7-10). I use 'prescriptive' here to refer to a subset of 'normative' covering what we ought to intentionally do and excluding evaluative claims (which, while normative, do not directly tell us what we ought to intentionally do, even *pro tanto*).

From this point of view, the prescriptive force of hypothetical agreement seems quite obvious. It relies on the claim that hypothetical choice tracks what we have most reason to do (and I take it as basic that 'what we have most reason do' has prescriptive force). There is, however, a second kind of worry. It is one thing to say that the fact that *I* would choose to do some thing justifies me doing the thing. It is quite another to claim that the fact that *someone else* would choose to do a thing justifies me doing anything. But, to the extent that I am not identical with the agent making a hypothetical choice to agree to something, there is a sense in which it isn't me agreeing; and, any idealisation will mean that the choosing agent and I are not identical.

This objection has some force, and I will have more to say about idealisation later.[7] Nevertheless, we cannot plausibly reject all forms of idealisation-based hypothetical agreement. For example, idealising me as perfectly rational, such that I make no errors in reasoning, does not undermine the prescriptive force of hypothetical choice for the real me. Denying this is incoherent – it would require us to say that I have more reason to act contrary to the way I have most reason to act; idealising me as perfectly rational does not change the reasons which apply to me, it simply makes me ideally responsive to those reasons which apply to my non-idealised self. Thus, the real question is which kinds of idealisation are appropriate in the context of a particular social contract theory. There is no cause to reject *all* idealisation-based hypothetical choice.

One might offer a different kind of objection to the idea of hypothetical agreement. My claim has been that hypothetical agreement simply tracks reasons. If that is right, hypothetical agreement looks like a spare wheel. If I am justified in doing something because I have a reason to do that thing, the fact that, suitably idealised, I would choose to do the thing adds nothing to the story. Hypothetical agreement drops out of the picture. It can then further be objected that talking about 'hypothetical contracts' is misleading, since it draws our attention to intuitions about sticking to

[7] The Interface, Chapter 5, Chapter 7.

agreements and keeping promises, rather than to whatever the relevant reason is which is justifying the hypothetical agreement.

This is a sensible worry, and in discussing hypothetical social contract theories it is important to realise that the hypothetical contract for the most part serves simply as a convenient metaphor.[8] Does that mean we should dispense with it, because it is just going to confuse matters? Well, no.

First, it is not an accident that we often connect 'agreement' with 'acting on the basis of one's reasons'. Consider a case in which the two might come apart: agreeing to something when drunk. In such a case, an agent's capacity to respond to her reasons has been impaired. The normative significance we intuitively attach to agreement depends on certain conditions being met.[9] When we look at some particular agreement, and ask if it is the kind of agreement which is normatively significant, one thing we are asking is whether or not the agreement was made in circumstances which have impaired the agent's capacity to respond to her reasons. And, one heuristic for answering that question is to ask, would a suitably idealised version of the agent have agreed?

Second, the contract metaphor is a particularly useful tool for *liberal* theories. Liberal theories have a special interest in freedom. However, like all states, the liberal state is coercive. Reconciling state coercion and freedom is one of the central concerns of liberalism. One popular way to do so is via some theory of consent or agreement. For reasons which need not detain us here, contemporary liberal theorists typically do not pursue reconciliations which are based on actual, historical agreement. But they nonetheless employ hypothetical agreement to validate state coercion. Agreement, in suitably idealised circumstances, and between suitably idealised

[8] For some suggestions as to what might be elided by that "for the most part" see D'Agostino, Gaus, and Thrasher's discussion of the relationship between hypothetical and actual agreement in contemporary social contract theories (2017: §3.1). See also Rawls, 1999a: 14-15; Stark, 2000; Chambers, 2010: 895-896.

[9] One could say that 'agreement' is only really *agreement* when the relevant conditions are met. This approach builds the normative content of my claim into the concept of agreement itself. My argument can be reframed, *mutatis mutandis*, to accommodate this point without any substantive change in meaning.

agents, is used by liberal theorists as a check that the political institutions they prescribe are consistent with appropriate respect for freedom.[10]

§1.3 Moral and prudential reasons

Social contract theories assume that, when making hypothetical choices, agents act on the basis of the reasons they have. However, different theories idealise their agents so as to be responsive to different kinds of reasons.[11] I will be discussing theories which rely on either moral reasons or prudential reasons (or both).[12]

We can get a fairly good understanding of what a prudential reason is quite quickly: We tell a child that she ought not to touch the hot stove. Why? Because it is hot, and when people touch hot things, they get burned, and getting burned is painful. When we tell the child that she ought not to touch the stove because it is hot, we are giving her a prudential reason to refrain from touching the stove. Avoiding the stove will advance one of her ends – avoiding pain – and she has a reason to advance her ends.[13]

Prudential reasons can be contrasted with moral reasons.[14] As before, we can get an

[10] I want to flag this point for careful consideration later. The way in which social contract theories model free agreement plays an important role in liberal accounts of political institutions. However, I will not be able to rely on this device in quite the same way as other accounts, and we will want to assess the extent to which this undermines my argument. See §4.2.

[11] The term 'reason' has several senses in contemporary philosophy. See Broome (2013: 46-47) or Lenman (2009) for helpful overviews.

[12] I do not mean to imply that these are the only kinds of reasons. One might, for example, think that aesthetic reasons constitute a further, distinct domain of normativity. I wish to remain neutral on such questions. Similarly, I am primarily concerned with what we have reasons to do, rather than what we have reasons to believe. As such, I do not discuss epistemic reasons, but I do not wish to deny they exist. Compare Broome, 2013: 25-29. In §2.4.1, *infra*, I consider the objection that prudential reasons are a kind of moral reason. See also n14, *infra*.

[13] Korsgaard, 1996: 36. Note I say prudential reasons track ends (or, equivalently, goals), not interests. My ends (or goals) and my interests need not align. I thank Sam Kiss for emphasising this point.

[14] If one is a moral contractarian (Gauthier, 1986; Gauthier, 1991), then insofar as one holds that there *are* moral reasons, one believes that all such reasons ultimately have a prudential basis; in that case, moral reasons are a subset of prudential reasons. However, this is not how we typically understand moral and prudential reasons, and I will mostly be using the terms in their more everyday senses, as referring to distinct domains of normativity.

initial handle on what a moral reason is with relative ease: We tell a child that she ought not to take a toy from the store without paying because that would be stealing, stealing is morally wrong, and she has a reason to avoid actions which are morally wrong.[15]

There are two further things to note. First, the distinction between moral and prudential reasons is a matter of the source and character of a reason's normativity, rather than the content of the reason's prescription for action. Thus, a moral reason and a prudential reason can have the same content, inasmuch as each prescribes that you act in exactly the same way, without being the same reason. Saying that an agent is responsive to prudential reasons implies only that the agent acts so as to rationally pursue whatever ends she happens to have. There is nothing to say these ends cannot be moral ends. In particular, we will not have said that the agent is selfish, self-interested, or amoral.[16]

The second thing of note is related to the context in which hypothetical agreement occurs. In social contract theories where the basis of agreement is moral reasons, it is typically the case that all of the agents in question have the same moral reasons.[17] In such a case, agreement between agents is a formality. Things are quite different when the basis of agreement is prudential reasons. Unless the agents in question just happen to share the same ends, they will each be responding to different, often conflicting, prudential reasons. In such a case, agreement is not merely a formality, and if we want the agents to agree, we will need some mechanism for resolving the conflict between them. This mechanism is provided by a theory of *bargaining*. There is typically no need for bargaining when the basis of agreement is moral reasons, but bargaining is central to an account based on prudential reasons.[18]

---

[15] Compare Rawls, 2001: 7; Moehler, 2018: 6-10, 12. I return to the distinction between moral and prudential reasons in §2.4, *infra*.
[16] Williams, 1981: 105; Gauthier, 1986: 7, 19; Gauthier, 1991: 17; Scheffler, 1994: 21-22; Barry, 1995: 32; Rawls, 1999a: 110-112; Rawls, 2005: 51; McCabe, 2010: 134; Abizadeh, 2011: 300-301. I discuss this further in §2.4 and §5.1.
[17] Ackerman, 1980: 327-328.
[18] Chapter 6.

At this point, I want to introduce some terminology to help mark the distinction between moral and prudential reasons. In the argument that follows, I will be using the term 'ought' in the context of normative claims underpinned by (among other things) either prudential or moral reasons. Many political philosophers will naturally be used to understanding 'ought' simply as '(morally-)ought', so it will be important to keep in mind that this is not the only sense in which I will use the term.

To avoid confusion, I will qualify my claims wherever possible to specify the kind of normative claim I am making. So, for example, consider the prescription, "You ought to refrain from stealing." This can be understood as a moral prescription: "You (morally-)ought to refrain from stealing." However, it can also be understood as a prudential prescription: "You (prudentially-)ought to refrain from stealing." When we understand the claim as a moral prescription, we presumably think of stealing as something that is morally wrong, and the moral wrongness of the act generates a moral reason (a duty) to refrain from performing the act. When we understand the claim as a prudential prescription, perhaps we have in mind the prospect of being caught and punished for stealing. Assuming that punishment is something we want to avoid, and assuming that we will be caught and punished if we steal, we have a prudential reason to refrain from stealing.[19]

All forms of prescription involve the provision of reasons.[20] The different varieties of prescription are distinguished by the different kinds of reason that are offered in support of the prescription being made. The fact that a prescription rests on a prudential reason, rather than a moral reason, doesn't make it any more or less

[19] In addition to '(morally-)ought' and '(prudentially-)ought', I will occasionally want to refer to a reason which could rely on either kind of normativity. I will denote this by '(in-some-sense-)ought'. Similarly, I will sometimes want to refer to reasons which might incorporate both moral and prudential dimensions. I will denote this by '(all-things-considered-)ought' (I expressly depart from Broome here (1999: 399-400; 2013: 25-26)). I will also want to say, from time to time, that we should not take such and such an objection seriously, or that we should accept such and such a conclusion. These claims refer to epistemic reasons. As I have said (n12, *supra*), I will not discuss epistemic reasons. I assume, without argument, that we (epistemically-)ought to believe true things. It is not obviously the case that, for example, we (prudentially-)ought to believe true things, though it is a plausible conjecture in most circumstances. For simplicity, I will just assume the conjecture. I will mark the distinction between theoretical and practical reason by reserving 'should' language for claims about theoretical reason and 'ought' language for claims about practical reason.

[20] Kramer, 1999: 80-92.

normative.[21] The child really does have a reason to refrain from touching the hot stove, and the reason in question is not a moral one. The prescription "You ought not to touch the stove" is justified by the child's prudential reason to avoid being burned. Since prudential reasons can justify prescriptions, they can in principle justify a prescription such as "You ought to establish and support liberal political institutions." And so, contractarian arguments, which rely on prudential prescriptions, can in principle justify the creation or maintenance of liberal political institutions.

To put this point another way, political philosophers sometimes have a tendency to use the terms 'moral' and 'normative' interchangeably, as if all normativity were a question of moral reasons. Everything I have to say relies on resisting this conflation. Normativity can be grounded in either prudential or moral reasons.[22]

§1.4 Categorising social contract theories

Political contractarianism is one of a number of possible positions falling under the rubric of hypothetical social contract theories. Prominent others can be labeled 'political contractualism' (and represented by, for example, John Rawls), 'moral contractualism' (represented by T. M. Scanlon), and 'moral contractarianism' (represented by David Gauthier).[23]

As this list suggests, social contract theories can be understood as varying on (at least) two dimensions.[24] The contractarian/contractualist axis describes the basis of agreement – the nature of the reasons to which the agents in the hypothetical choice situation respond.[25] Contractarian theories exclude all moral reasons, and allow only

[21] Broome, 2005a: 324-325; Broome, 2005b: 3.
[22] Moehler, 2018: 8-9.
[23] Rawls, 1999a; Rawls, 2005; Scanlon, 1998; Gauthier, 1986; Gauthier, 1991.
[24] See also Hampton (1991: 32-33) and Rossi (2010: 27-28), though each draws these lines in a slightly different way.
[25] The term 'contractarian' has sometimes been used to refer to social contract theories in general. In his earlier work, Rawls (1999a: 14-15) describes his method in this way. This usage has become somewhat outdated, however. I will be using the term 'contractarian' in its more specialised sense.

prudential reasons to serve as a basis for hypothetical agreement.[26] Contractualist theories, on the other hand, use at least some moral reasons in the construction of the agreement.[27] The prescriptions of Rawls's theory, for example, are ultimately based on the claim that we have a moral reason to treat our fellow citizens as free and equal.[28]

The political/moral axis defines the object of agreement – the scope of application of the social contract theory in question. Moral contract theorists (Scanlon and Gauthier) derive the content of morality as a whole from a conception of a hypothetical contract. Political contract theorists (such as Rawls) limit the scope of their hypothetical contract to the 'political' domain (although there is disagreement about how this domain is delineated).[29]

Moral contractualism takes the contrary position to political contractarianism on each of the two dimensions. It is engaged in a wholly distinct project, and so I won't discuss it further. On the other hand, moral contractarianism and political contractualism each share one dimension with political contractarianism, and differ on the other. Contrasting political contractarianism with each of these theories will be instructive. In §1.5, I discuss the relationship between political contractarianism and moral contractarianism. In Chapter 3, I discuss the relationship between political contractarianism and political contractualism.

See Gauthier, 1997: 132-135; Darwall, 2003b: 1-8; Cohen, 2008: 337-343; Cudd and Eftekhari, 2017; D'Agostino, Gaus, and Thrasher, 2017; Weale, 2013: xii; Moehler, 2018: ix, 11-12. Matravers (2000: 98 n1) offers an alternative terminology.

[26] Strictly, this claim is too strong. As I will discuss (§2.3.3, particularly n42), there is a sense in which moral reasons can serve as a basis for agreement in political contractarian theories: such theories need not exclude as reasons moral propositions about which there is, in fact, no dispute (for a given population of agents). For the time being, I neglect this complication. See also the discussion of outcome-oriented moral reasons in §2.4.3.

[27] Moehler, 2018: 6-7, 12.

[28] See Rawls's contrast between an overlapping consensus and a *modus vivendi* (2005: 168).

[29] Compare Vanderschraaf, 2019: 190-191.

§1.5 Contractarianism and moral realism

As set out by Gauthier, contractarianism is explicitly intended to be an alternative to traditional moral realism.[30] Gauthier's starting point is the assertion that we cannot plausibly continue to believe that the traditional claims of morality are true. We may once have seen the universe as "purposively ordered". But as soon as we discard the childish notion that some divine lawgiver has created a moral code which we must follow, or that there is some special purpose for which we exist, we are left with nothing but the cold realisation that we are accidents of organic chemistry living in a universe which doesn't care what we do or how we treat one another.[31] Morality, traditionally conceived, is a delusion (or perhaps a pretension) from which we have awoken.

On Gauthier's account, the theistic and teleological foundations of morality have been discredited, and morality can no longer rely on them for its prescriptive force. Moral contractarianism is an attempt to refound morality. It aims to give traditional moral propositions an alternative form of prescriptive force, by showing how they are ultimately grounded in prudential rationality. It is moral philosophy for grown-ups, shorn of mysticism and make-believe.[32]

Accepting moral contractarianism means rejecting moral realism. It means accepting that there really isn't anything wrong with torturing a child for fun. If torture would amuse you, or would help advance one of your ends, then (all else equal) you lack

[30] 1986; 1988; 1991. In addition to Gauthier, Moehler (2018: 12) identifies Hobbes (1996), Hampton (1986), and Kavka (1986) as moral contractarians. To these, I would add (see n45, *infra*) Moehler (2018) and Vanderschraaf (2019), and perhaps Muldoon (2016). Note that I say moral contractarianism entails a rejection of *traditional* moral realism to allow for the possibility that a moral contractarian might view her theory as a variety of moral realism. Moral contractarianism is a cognitivist theory – its prescriptions are truth apt – and it further claims that its prescriptions are true. So long as it claims that its prescriptions are genuinely *moral* prescriptions, as Gauthier does (Morris, 2013: 595-596, 598), then moral contractarianism seems strictly to meet the criteria defining moral realism. Nonetheless, it is radically at odds with what most people have in mind when they think of moral realism. In the text subsequently, when I refer to moral realism, I will mean traditional moral realism. See n14, *supra*. Compare Moehler, 2018: 6-7, 12.

[31] Gauthier, 1991: 15-16.

[32] Gauthier, 1988: 385; Matravers, 2000: 162-163; Rawls, 2005: 51-53; Thoma, 2015: 3337. Compare Korsgaard, 1996: 18.

a reason to refrain from it.

For many people, this will be difficult to accept. Nonetheless, a moral contractarian will argue that we simply have no epistemological basis for thinking that there are such things as moral reasons. Moral properties do not figure in our 'best explanations' of the phenomena we observe; they are problematically 'queer' properties.[33]

Given that moral realism and moral contractarianism are incompatible, the success of Gauthier's argument ultimately depends on one's attachment to moral realism. I am not convinced that I can sincerely reject realism. As Ronald Dworkin puts it, "any reason we think we have for abandoning a conviction is itself just another conviction... [It] is startlingly counterintuitive to think there is nothing wrong with genocide or slavery or torturing a baby for fun. I would need very powerful, indeed unanswerable, reasons for accepting this..."[34] This is to say, any argument which we can make against moral realism will, in the end, rest on an intuition of some kind, whether it is an intuition about the rules of inference, or about what is epistemically reasonable, or whatever. Our confidence in that foundation will have to be sufficiently great to defeat our conviction that genocide, slavery, and torture are wrong. Otherwise, moral realism will survive the process of reflection, and whatever epistemological or ontological intuition it was which threatened moral realism will have to be abandoned instead. We "cannot think any [sceptical] argument a decisive refutation of a belief it does not even dent. In the beginning, and in the end, is the conviction."[35] I find myself less confident in moral realism than Dworkin appears to be, and more worried that it might, in the end, have to be set aside as indefensible. But, for the time being, when pressed to choose, I will still insist that there are some things which are 'really' wrong. That being so, I reject moral contractarianism.

---

[33] Gauthier, 1991: 15-16; also 1986: 54-58; Narveson 1988: 110-115. On the argument from best explanation see Harman (1977: 3-10); on the argument from queerness see Mackie (1977: 38-42).
[34] Dworkin, 1996: 118.
[35] *Ibid*. Narveson (1988: 115-118), in contrast, argues that we should resist the conflation of moral (or 'metaphysical') intuitionism and what he calls 'methodological intuitionism'. I doubt a hard distinction between the two can be maintained, but in §2.3.2 I will rely on the claim that, as a matter of fact, there is likely to be less disagreement about methodological claims than about fundamental moral claims.

Nevertheless, rejecting moral contractarianism does not entail rejecting political contractarianism. Political contractarianism is engaged in a different kind of project, and concerned with a different kind of problem. It starts from a conception of what politics is. Politics, on my account, is about getting other people to do stuff. People are entities with goals and ambitions – with *preferences* over possible states of affairs.[36] They exist in a world in which they often cannot achieve their various ends without the coordination of other people.[37] 'Coordination' is here understood to include both positive actions you might take to advance my ends and non-interference on your part in actions I (or third parties) might take to advance my ends. Other people can have reasons, either prudential or moral, to coordinate with me, but might also have reasons not to coordinate with me. Politics occurs when one agent attempts to get another agent to coordinate by offering her a reason to do so.[38]

Politics, then, involves agents offering one another reasons. However, a problem sometimes arises when agents offer one another moral reasons. One of the features of the world we inhabit is the fact of pluralism.[39] There is disagreement about values, about the truth of particular moral propositions, and about the truth-status of moral

[36] For discussion of the concept of a preference, see §5.1. Briefly, to say that *A* has a preference for a state of the world, S_1, over another state of the world, S_2, is to say that *A* believes herself to have greater reason to act in such a way as to realise S_1 than S_2.

[37] That is, people exist in the 'circumstances of justice' (Rawls, 1999a: 109-112; Hubin, 1979: 19-24). For further discussion, see §5.5, *infra*. Note that I refer to agents *coordinating* with one another, rather than *cooperating* with one another. As I discuss in more detail later (§2.1, §4.2, §5.5, §6.1, and §6.6), I want to leave open the possibility that coordination is coerced, and the more commonly used term – 'cooperation' – risks giving the impression that the agents' coordination is uncoerced (Rawls, 2001: 6; compare Vallier, 2019: 115).

[38] For a more detailed discussion, see §2.1.

[39] Rawls, 2005: 36-37. Rawls believes that pluralism (more specifically, 'reasonable pluralism') "is a permanent feature of the public culture of democracy" and that "a continuing shared understanding on one comprehensive religious, philosophical, or moral doctrine can be maintained only by the oppressive use of state power." (For support for this view, see the discussion of the burdens of judgement, *ibid*.: 54-58.) It is sufficient for my purposes if the fact of pluralism is an accurate description of the circumstances in which many of us presently live, whether or not it is inevitable given liberal political institutions (contrast Moehler, 2018: 1). Clearly, the fact of reasonable pluralism presupposes the fact of pluralism *per se*; Rawls assumes that, in addition to a variety of reasonable comprehensive doctrines, there will always exist a variety of unreasonable comprehensive doctrines (Rawls, 2005: 39; 63-64). I do not propose to argue for the fact of pluralism; my argument should simply be understood as contingent on the existence of pluralism.

propositions in general. Suppose Angela were to suggest a particular joint course of action to Brian, and were to support her proposal by claiming that pursuing the course of action would maximise utility. If Brian is a utilitarian, then Angela's claim would give him a reason to coordinate his actions with her. If Brian is not a utilitarian, however, the fact that the course of action proposed by Angela would maximise utility simply doesn't seem to provide him with any reason to act as she suggests.[40]

The existence of moral disagreement means that moral reasons will sometimes, and perhaps often, fail to motivate coordination when offered by one agent to another. Since politics is (by definition) about getting other people to coordinate, it must be composed of reasons that are *effective* in motivating coordination. To the extent that moral reasons are not motivating, they are not the appropriate constituents of politics. Prudential reasons, however, don't suffer from the same problem. If Angela suggests to Brian that he should coordinate with her because doing so is a means of accomplishing one of his goals, then so long as she is correct her claim will always be motivating. It will always give Brian a (*pro tanto*) reason to coordinate.

Political contractarianism, then, is a theory of politics in which agents employ prudential reasons rather than moral reasons for the purpose of getting others to coordinate with them. But the justification for excluding moral reasons is not grounded in the general moral scepticism which motivates moral contractarianism. Political contractarianism is instead motivated by the fact of pluralism, and by the consequent problem of moral disagreement. In the case of Angela and Brian, we can stipulate that Brian is a deontologist, and Angela a utilitarian. The problem of moral disagreement is present, and the solution offered by the appeal to prudential reasons remains valid, regardless of the fact that each participant is a moral realist.

Political contractarianism, in recommending the use of prudential reasons in politics where the problem of moral disagreement arises, is not committed to either

[40] For a more extensive discussion, see §2.3.3.

endorsing or denying moral realism. Political contractarianism is as much a theory designed to accommodate the *competing* moral realisms of co-citizens as it is a theory which can give motivationally effective reasons for action to moral sceptics.[41] It is a theory about how we can justify ourselves to other people who do not share our moral convictions; it does not require that we reject our own moral convictions as unjustified. In this sense, it is a prototypically *liberal* theory.

It should be clear, then, that there is no contradiction between political contractarianism and moral realism. Neither, of course, is there a contradiction between political contractarianism and moral contractarianism. While I (most days) endorse moral realism, and have criticised moral contractarianism for its inconsistency with moral realism, the political contractarianism I am defending is strictly neutral between moral realism and its competitors.[42]

§1.6 Contemporary contractarianisms

Moral contractarianism and political contractarianism are distinguished by what they take to be the object of contractarian agreement. Moral contractarians aim to derive the content of morality as a whole. Political contractarian theories have a narrower scope. Political contractarians, broadly speaking, aim to identify a set of normative principles to regulate political interaction. In this section, I begin by examining Gauthier's own turn from the moral to the political.[43] Doing so provides a useful template for analysing political contractarian theories and clarifies some key questions with which such theories must contend. I will then consider three more recent accounts – by Michael Moehler, Peter Vanderschraaf, and Ryan Muldoon.[44]

[41] Compare Barry, 1995: 39; Moehler, 2019: 26-27.
[42] Compare Gauthier, 1997: 142.
[43] 1997.
[44] Muldoon, 2016; Moehler, 2018; Vanderschraaf, 2019. A number of other authors have made important contributions to the literature on contractarianism. Most influential among these, whether as critics or proponents, have been Buchanan (1975), Narveson (1988), Binmore (1994, 1998), and Gaus (2011). I will not have the space to consider these accounts in detail here (though I will draw on them where necessary). My hope, in focusing on the most recent literature, is that it will incorporate the most important insights of the earlier contributions.

These accounts can sensibly be understood as contractarian (if not necessarily political). Discussing them gives me an opportunity to examine the current state of the contractarian literature. I will note the continuities between these approaches and mine, and also identify the key points at which I will depart from the existing accounts. Where aspects of this discussion unavoidably presuppose a familiarity with some of the more technical elements of contractarian theories, I will note the later sections of my argument which provide a more systematic introduction to the ideas in question.

The three recent accounts I will examine are probably better understood as examples of moral contractarianism.[45] Most recent work on these questions does not draw as sharp a line as I do between moral and political theories. In part, this is a result of the rather flexible boundaries of both 'moral' and 'political'. If one has certain kinds of value commitments (social libertarian, subjectivist, say), there are simply no moral questions about actions which do not affect other people – such actions are amoral (in the strict sense), so all is permitted. Similarly, 'political' can be understood in a very broad way – as equivalent to 'social' – such that it covers all interactions between people. If one has *both* of these inclinations, the distinction between 'moral' and 'political' may appear redundant.

We should resist this tendency. First, because not everyone is going to share the narrow conception of 'moral' upon which it relies.[46] Second, however, because the moral-political distinction is not *only* tracking a range of actions. Moral contractarians typically take their task to be providing a contractarian basis for 'morality'.[47] Even moral contractarians who are willing to be strongly revisionist

[45] This is true of Moehler's account, for instance (2018: x, 10, 12, 20, 21, 23, 50-51, 95, 105-106, 108, 133-139, 160). It is plausibly true also of Vanderschraaf's account (2019: 226-236, 271-273, 275-276, 280-283, 286, 304-305, 308-311, 318-323). Muldoon, in addition, might not even be a contractarian – in places he seems more appropriately read as a kind of moral minimalist (2016: Chapter 1, §2.2, §3.1, §3.2, §3.3, §3.4.3, §4.5.1, §4.6, §5.1, §5.3.2, §5.5, §5.6, Chapter 6; for my discussion of moral minimalism, see the Introduction, *supra*).
[46] Compare Mill's discussion of 'social rights' in Chapter IV of *On Liberty* (1974: 156-158). Similarly, if one's future self or supernatural beings can count as 'people', the scope of 'affects other people' is *de facto* unlimited.
[47] Gauthier, 1991: 15-22.

about the traditional content of morality, such as Gauthier, accept that there are limits to how far one can stretch the concept of 'morality' before one must simply concede that one has left 'morality' behind and is now talking about something conceptually distinct. For Gauthier, the idea of impartiality must play *some* role in an account of something that is genuinely *moral*.[48] This means that some degree of congruence between morality, traditionally understood, and the conclusions of a contractarian theory is taken to be a success condition of the theory in question. To the extent a moral contractarian cannot show that (more or less) traditional moral principles could be the outcome of a rational bargain between the parties to a social contract, she has *failed*.[49]

The same is not true for political contractarian accounts.[50] Strictly, it is not a failure (though it may be a tragedy), in a political contractarian account, that the theory prescribes morally outrageous things. As it happens, I will argue later that, in the kinds of circumstances in which citizens of contemporary societies tend to find themselves, political contractarian accounts should endorse political institutions that quite closely resemble those of liberal democratic societies, and so won't typically endorse morally outrageous conclusions. However, this doesn't mean the distinction between moral and political theories is unimportant. Under more extreme conditions, the prescriptions of the theories will probably come apart. Equally, moral and political contractarian theories are likely to diverge on the question of inclusion (of who gets represented in the hypothetical choice situation).[51]

Despite the fact that the accounts of Moehler, Vanderschraaf, and Muldoon are not straightforwardly political contractarianisms, it will be useful to examine them here. These accounts present highly sophisticated discussions of many of the problems in which I am interested. To a significant extent, the value of these discussions is

[48] Gauthier, 1986: 4-6, 16-17, 233-235, 266-267; Thoma, 2015: 3335-3337; Moehler, 2018: 10 n33; Vanderschraaf, 2019: 274, 286.
[49] Moehler, 2018: 10. Compare Vanderschraaf, 2019: 226-233, 273, 308-309, 318-319, 321-322.
[50] At least, not necessarily. A political contractarian may take congruence with traditional moral principles to be a success condition for his account. But this requirement isn't 'built in' for political contractarian theories as it is for moral contractarian theories.
[51] I discuss questions of inclusion and exclusion in §5.5.2 (particularly n66) and §8.5.

independent of the author's larger metaethical commitments (or ambiguities). To put that another way, even if these authors would not understand themselves as political contractarians in my sense, it is relatively straightforward to see how political versions of their accounts would run, so the comparisons will still be instructive.

§1.6.1 Gauthier's agnostic contractarianism

The most straightforward place to begin is with Gauthier's own account of "Political Contractarianism".[52] Unlike some other contractarian accounts, Gauthier's political version of contractarianism is 'agnostic' – it is not simply an extension of his moral contractarianism (though they are, of course, compatible).[53] On Gauthier's account, society is understood as a scheme for mutual advantage; its purpose is to give us access to goods which we could not realise without the coordination of other people.[54] Coordination requires interaction between agents, and interaction is to be governed by certain principles. The point of political contractarianism, on Gauthier's account, is to develop "a mode of justification for norms governing interaction", specifically in circumstances where we "may neither expect nor require [our] fellows to share [our] own orientation to values and norms."[55] Because of disagreement about values and norms, and our limited powers, the norms which are to govern interaction must be justified by appealing only to "the norms of deliberative rationality... All other norms, deriving, say, from a religion, or from a view of the world as characterized by objective value, are ones that only fellow believers in the religion or world view can recognize as ones that they themselves would accept."[56] The norms which govern interaction neither affirm nor deny the particular beliefs

[52] 1997. This article has attracted surprisingly little critical attention.
[53] Examples of accounts which defend a political contractarianism *on the basis of* moral contractarianism include those developed by Narveson (1988) and Binmore (1994, 1998), and arguably Moehler (2018) and Vanderschraaf (2019), to the extent we interpret Moehler and Vanderschraaf as political. Gauthier would describe these as 'atheistic' contractarianisms (1997: 142).
[54] 1997: 135.
[55] *Ibid*.: 132.
[56] *Ibid*.: 133. By "the norms of deliberative rationality" Gauthier means prudential reasons (1991: 19; 1997: 132-133).

about value which citizens hold, but these beliefs about value "are regarded as unsuitable for the purposes of *social* justification..."[57]

Although Gauthier and I differ somewhat in our vocabulary, I accept his basic framing of the problem, including metaethical agnosticism, the fact of pluralism, consequent constraints on public reason, and the appeal to prudential reasons. However, I note three main points of divergence between our accounts.

First, in line with many contract theorists, Gauthier seems to accept a certain kind of public justification principle.[58] This appears to commit him to the idea that the reasons offered in support of political institutions must be of a kind any citizen could (freely) accept. But why is this important? There are two distinct issues here. One we can describe as the *exclusion problem*. From the point of view of political contractarianism, my motivation for offering you a reason you can accept is supposed to be that doing so is the only way in which I can secure your coordination. But this assumes that I am interested in coordinating with you.[59] While it is perfectly plausible to think that I am interested in coordinating with some people in order to advance my ends, why do the reasons I offer in support of political institutions have to be acceptable to everyone, as opposed to being acceptable simply to the people with whom I am interested in coordinating?[60] The second issue we can describe as the *coercive baseline problem*. I may be able to persuade you to cooperate with me by offering you a reason you can (freely) accept, but why would I bother if I can

[57] *Ibid*.: 142 (emphasis added).
[58] *Ibid*.: 133. He writes, "[Moral reasons] may not then be appealed to directly in the defence of social practices *that must be justifiable to each*." (Emphasis added.) For a discussion of the public justification principle, see Gaus (2011: 263ff), Vallier (2014: 24-31).
[59] It assumes I have a disposition to be *reasonable*, in Vallier's terms (2014: 147-148). See the discussion of Vallier's account of idealisation in the Interface, *infra*.
[60] Gauthier's discussion of including children among the contracting parties constitutes one aspect of the exclusion problem (1997: 146-148). In "Political Contractarianism", Gauthier adopts a 'whole life' perspective when characterising the agents participating in the hypothetical social contract, so that children are represented in contractarian bargaining by rational proxies aiming to secure for them the means of growing into their full rational capacities safely and securely. This raises a number of difficult questions for a contractarian account, not least because the children represented in bargaining will not have determinate conceptions of the good. There is also the question of children not yet born, and of future generations more generally. I doubt that a contractarian theory can consistently pursue this method of including children among the contracting parties, but see Vanderschraaf (2019: 287-292, 301-304) for further discussion.

simply coerce you? It is your compliance with my demands that I need, not your approval. The standard moral response to these questions – that persons have an intrinsic value which gives us decisive moral reasons to treat them with equal concern and respect – is not available to the contractarian. I depart from Gauthier in my response to these questions.[61]

Second, Gauthier and I differ in the extent to which we think revisions to existing political institutions follow from a contractarian argument. Gauthier is concerned that the "idea of agreement among hypothetical deliberators choosing their terms of interaction may seem too indefinite to afford any direct conclusions about the justifiability of actual social norms and practices. There is, after all, no social arrangement that would seem best for everyone, so would not any proposed norms be rejected by some?"[62] He goes on to reply that a contractarian account does not set out to give every participant in political institutions her most preferred outcome, but rather to select an outcome that (i) is preferred by all participants to the absence of cooperation, and (ii) "ensures that the proportion between benefits and the contribution that each [participant] chooses to make be roughly the same for all [participants]."[63] This second condition can be expressed more generally as a requirement that the outcome selected represent a rational compromise between the ambitions of each participant, as determined by the contractarian account's particular bargaining theory.

Gauthier nonetheless thinks that these conditions are likely to underdetermine the political institutions to which citizens have most reason to agree, presumably because he does not believe that the relevant bargaining theory will select a *unique* outcome: "[The] possibility that other norms, practices and institutions might be equally reasonable to accept is no argument against an existing social order. Only if the existing society could not plausibly be represented as the outcome of reasonable agreement on terms of interaction would there be grounds to reject it..."[64] He

[61] For my discussion of the baseline problem, see §7.5. For my discussion of inclusion, see §8.5.
[62] 1997: 140.
[63] *Ibid*.
[64] *Ibid*. Gauthier doesn't give us a clear statement of how we are supposed to understand 'reasonable'

concludes that the "task of contractarian justification is made *sufficiently* determinate by appealing to and only to the deliberative rationality of the members of society."[65] I take his point to be that the appeal to prudential reasons is sufficient for the purpose he has identified – working out if an existing social order could plausibly be represented as the outcome of contractarian agreement, and therefore if said order can reasonably be rejected.

Gauthier's concerns about underdetermination and uniqueness prefigure a debate that continues in the contemporary literature.[66] This debate has a couple of aspects it will be helpful to distinguish. One involves the degree of determinacy to which political contractarian theories aspire. As just set out, Gauthier's account is relatively expansive – he anticipates a relatively wide range of political arrangements satisfying his conditions.[67] As a matter of practical application, then, his account is ('small-c') conservative, in that we can imagine that many more sets of existing political institutions will be justified on Gauthier's account than on a less expansive account (and in this, we might add, Gauthier is simply being a good Hobbesian). I will ultimately argue for a more narrowly drawn set of contractarian conditions – and consequently greater determinacy and less deference to the *status quo* – but I would not wish to argue that a contractarian account can be *fully* determinate.[68]

The other aspect of the uniqueness debate we should note is what is sometimes called the equilibrium selection problem.[69] If the rest of one's account does not uniquely determine a set of political institutions for a particular society, there is a question about how to select one set of political institutions from among the alternatives one's other criteria have not excluded. As noted, in the event that

here. See also the discussion in Moehler (2018: 59-66).

[65] 1997: 140 (emphasis added).

[66] Binmore, writing around the same time, raises similar concerns (1994: 57-58; 1998: 4-5). Prominent recent discussions include, for example, Gaus (2011: 294ff), Thrasher (2014), Thoma (2015), Muldoon (2016: Chapter 6), Vallier (2016: 611-612), Moehler (2018: 4-5, 109-110, 121-129, 140-158, 161), and Vanderschraaf (2019: 24, 58-59, 116-117, 165-175, 188, 223, 301, 308-311, 318, 320).

[67] Compare Gaus's discussion of the socially eligible set (2011: 321-332); or Vanderschraaf's discussion of the justification of non-democratic regimes (2019: 222-223).

[68] See §3.2.1, and n27 in particular; and §7.8.

[69] Vanderschraaf, 2019: 24.

multiple sets of institutions survive contractarian scrutiny, Gauthier defaults to *existing* institutions. This is a position broadly followed by a set of theorists we can identify as *conventionalists*.[70] Contemporary conventionalist accounts are typically underpinned by evolutionary game theory.[71] They argue that a set of institutions for a particular society is *justified by the fact that* it is the set that has emerged from a particular process of historical development as a game-theoretically stable solution to the society's coordination problems.[72]

One key question for conventionalist theories is whether they can provide a genuinely normative account of political institutions, as opposed to a merely explanatory account.[73] Can conventionalists give us reasons to endorse particular institutions, as opposed to explaining how we have come to have those institutions (even though we could have had, or could still have, others)? This question is linked to the question of the degree of determinacy in the following way. To the extent that an account is fully determinate, there is only one set of institutions uniquely justified. Any steps away from full determinacy open up an equilibrium selection problem; and, to reiterate, at least some departures from full determinacy seem inevitable. The more expansive the range of eligible institutions, the less plausible the claim that any particular set of institutions is justified simply on the basis that it has emerged from a historically contingent process. The worry is that any reliance on historical contingency makes the resulting institutions *arbitrary*, and hence unjustified – for example, to those agents who would be better off under alternative elements of the eligible set.[74] A set of institutions can be game-theoretically stable, in the sense that no individual agent has an incentive to *unilaterally* deviate from them, while

[70] Moehler (2018: 12) and Vanderschraaf (2019: 49-53) identify this tradition most strongly with Hume. Among contemporary authors, Moehler (2018: 12, 154-155) picks out Sugden, Binmore, Skyrms, Alexander, Vanderschraaf, and arguably Muldoon, as conventionalists. There are also conventionalist elements in Gaus (2011: 409-417), but there the picture is more complicated (Moehler, 2018: 148-153, 155-158). I do not mean to imply here that all conventionalist approaches are as forgiving (or as restrictive) as Gauthier's in "Political Contractarianism". Much depends on how stringent the other requirements of the particular account are – which equilibria are ruled in or out.
[71] Vanderschraaf, 2019: 191. The seminal text here is by Skyrms (1996).
[72] See, for example, Gaus (2018). For an excellent introduction to pure and partial coordination games, see Vanderschraaf (2019: 24-25, 44-84).
[73] On this distinction, see Vanderschraaf (2019: 191, 215).
[74] Moehler (2018: 154-155) offers an objection of this kind against Muldoon (2016), for instance.

nonetheless being unjustified when considered from the point of view of what agents would agree if they were bargaining *ex ante*.[75] Perhaps we would worry about this less if the set of eligible options fell within a tightly bounded range. But it is by no means clear that this will be the case, particularly for contractarian accounts, which cannot rely on even minimal moral commitments to circumscribe the eligible range.[76] As I have already noted, in the argument to follow I will defend a more determinate account of political contractarianism than Gauthier. I will also reject versions of political contractarianism that are excessively arbitrary, in the sense that they do not adequately justify their selection of political institutions to the relevant constituency of agents.[77]

The third point of divergence between my account of political contractarianism and Gauthier's is a consequence of the first two. Because of the different approaches we take to the problems of exclusion and the coercive baseline, and to degrees of determinacy and the perspective from which institutions must be justified, Gauthier and I ultimately recommend substantively different political institutions. In "Political Contractarianism", Gauthier sketches what we might tentatively describe as a sufficientarian principle of distributive justice for a society organised according to his theory.[78] In contrast, my account will be broadly egalitarian.

[75] See §5.6. Of course, we are not always solely interested in an *ex ante* perspective. For example, given costs associated with transitioning from one set of institutions to another, it may be that we have reason to remain where we are. In that event, however, the explanatory story of how we got where we are plays no role in the justification of existing institutions – the justificatory weight is carried by the assessment of the costs of transitioning. See Gaus (2011: 413-415) for a more robust argument for the normative significance of contingent history. Similarly, the evolution of a social contract over time is one of the principal concerns of Muldoon's work (2016). I thank David Miller for helpful comments on this issue.

[76] See the discussion of moral minimalism in the Introduction.

[77] See §3.2.1 n27, §5.6, and the discussion of exogenous determination in Chapter 8.

[78] 1997: 138-139. That the description is tentative is due both to the fact that Gauthier's outline is quick and schematic, and to the fact that while 'sufficientarian' seems to me to roughly capture his intention, the fit is by no means perfect. For an argument to a similarly 'sufficientarian' conclusion, although with quite different reasoning, see Moehler (2018: 18, 114-139).

§1.6.2 Moehler's two-level contractarianism

Having examined an early sketch of political contractarianism, let me now turn to a set of more recent authors working in a broadly similar theoretical space. Consider first the account presented by Michael Moehler.[79] I have already noted that Moehler should strictly be understood as a moral contractarian. The key fact with which Moehler begins is deep and persistent pluralism, and especially pluralism about morality:

> "In morally pluralistic societies, agents may hold irreconcilable moral ideals, or irreconcilable interpretations of such ideals, and as a result may hold different views about the demands of justice… If society enforces one such particular view of justice that relies on substantive moral ideals, then doing so may lead to social unrest… In modern pluralistic societies, especially under the assumption of deep moral pluralism, the ideal of a fully just society as judged from the perspectives of all members of society is unattainable…"[80]

Clearly, my project has much in common with Moehler's. We each begin with the fact of pervasive and persistent moral disagreement, and with a concern that this makes the implementation of an uncompromised conception of justice practically impossible.

From this starting point, Moehler aims to 'complete' the programme Gauthier articulates in *Morals By Agreement*.[81] He charges Gauthier with two heterodoxies, one against contractarian methodology, the other against rational choice theory. Gauthier departs from contractarian orthodoxy, Moehler thinks, in allowing a variety of moralised elements to enter his theory.[82] He departs from orthodox rational choice theory in how he develops his account of compliance and in his account of

[79] 2018.
[80] *Ibid*.: 1. See also *ibid*.: 5-6.
[81] *Ibid*.: 23.
[82] *Ibid*.: 23, 51-54, 56, 58-60, 61-62, 64-65. He makes a similar critique of McClennen (Moehler, 2018: 62-64).

bargaining.[83] Moehler aims to (i) argue for Gauthier's conclusion – that something recognisable as 'morality' can be derived from prudential reason – while (ii) avoiding any lapse into moral minimalism and (iii) remaining strictly orthodox in his account of rationality. Given that I am not concerned with moral contractarianism, I will set aside the first of these constraints. Nonetheless, a political contractarian theory should pay close attention to each of the other two. I endorse the requirement that contractarian theories avoid any moral commitments.[84] However, I will not endorse Moehler's rational choice orthodoxy without reservation.[85]

Moehler's proposed solution to the problem of deep pluralism relies on a distinction between two 'levels' of moral theory. The first level is constituted by 'traditional morality'.[86] This is the set of ethical and metaethical beliefs held by particular agents, including the foundational moral commitments according to which they evaluate possible actions and states of affairs. It provides the basic premises which are available for (traditional, first-level) moral justification by the lights of a particular agent.

When agents share sufficient moral premises, they can resolve disagreements by appeal to first-level morality. However, under conditions of deep pluralism, certain cases will arise in which there is a disagreement between agents about some proposed course of action or state of affairs, but where the agents in question disagree so deeply about morality (traditionally understood) that no intersubjective justification is possible between them, despite their best good-faith efforts.[87] In some subset of these cases, agents will attach extreme importance to resolving the disputes in their own favour. So much so that, in these cases, they will be willing to act in ways which threaten the lives of other agents (or of people close to them). Moehler supposes that such threatening or destructive actions are costly for agents, among

[83] *Ibid*.: 23, 51-58, 60-61, 62, 104-107.
[84] With one exception, specified in §2.3.3 (particularly n42). Compare Moehler's discussion of the principle of subsidiarity (2018: 163-165).
[85] In Chapter 8, I propose a revision of orthodox rational choice theory as it is applied to the theory of coalitions.
[86] 2018: 6-7. Compare my discussion of 'traditional' moral realism in §1.5 (particularly n30), *supra*.
[87] 2018: 15. Compare my discussion of Quong's impossibility problem in §3.2.1, *infra*.

other reasons because they "endanger the benefits of peaceful long-term cooperation…"[88] It is cases with these features which are Moehler's central concern. He calls them 'cases of conflict (in the strict sense defined)'.[89]

The second level of moral theory is constituted by what Moehler calls the 'purely instrumental approach to morality' or just 'pure instrumental morality'.[90] As Moehler defines it, pure instrumental morality contains no (traditional) moral premises, "but only an overarching end that the agents aim to reach..."[91] Moehler argues that, in cases of conflict (in the strict sense defined), agents have an incentive to avoid costly destructive actions by finding a principle of conflict resolution under which each party will do better than by engaging in conflict. By definition, in cases of conflict (in the strict sense defined), first-level morality cannot resolve the conflict. The role of second-level morality is to provide a principle of conflict resolution for cases of conflict (in the strict sense defined).[92] "The only basis for the justification of a principle of conflict resolution for cases of conflict in the strict sense defined is that agents are rational, they share a similar human nature and empirical environment, and they have an overarching interest in securing peaceful long-term cooperation."[93] From these materials, Moehler constructs a model of a rational agent ('*homo prudens*') and a stylised environment ('empathetic contractor theory') in which agents so characterised must select a principle to regulate their conduct in cases of conflict (in the strict sense defined).[94] Crucially, he argues that the elements of his model are "idealized but empirically defensible and, according to the traditional

[88] 2018: 15. Compare my discussion of the costs of conflict in §5.5, *infra*.
[89] 2018: 15-16, 99-100.
[90] *Ibid*.: 7-10, 95-139.
[91] *Ibid*.: 8. Importantly, this does not amount to saying that it includes no normative premises (*ibid*.: 8-9). Compare §1.3, *supra*. Moehler's phrasing is ambiguous, here, over the claim that *each* agent has a *particular* overarching end, and the claim that the agents collectively *share* an overarching end. It is tempting to read him as intending the former – that is more in keeping with his general contractarian orientation – but he actually seems to intend the latter. See *op. cit*.: 17, 101; and the text, *infra*.
[92] 2018: 16-17.
[93] *Ibid*.: 17. See n91, *supra*. One problem with assuming that agents share an overarching interest in securing peaceful long-term cooperation is that it elides the possibility that not *all* agents have an interest in securing cooperation with *all* other agents (or that agents might give differing priority to such a putative interest, such that for any particular agent the interest might not be overarching). See the discussion of the exclusion problem in §1.6.1, *supra*, and in §5.5.2, *infra*.
[94] 2018: 17, 95-114.

understanding of morality, nonmoralized…"[95] This allows him to avoid Gauthier's slide from contractarianism to moralised contractualism. He goes on to argue that, given this model, rational agents would select what he calls the 'weak principle of universalization' as their principle of conflict resolution.[96]

I accept much of Moehler's framing and argument here.[97] I am similarly concerned with deep pluralism. His cases of conflict (in the strict sense defined) are substantially similar to what I have characterised as cases of highest-order disagreement about justice. We each identify instrumental rationality as the only plausible source for a principle of coordination under conditions of profound moral disagreement.[98] We each develop a contractarian model on that basis, grounded in empirically defensible, non-moralised idealisation.[99] However, I also depart from Moehler at a number of points.[100] I focus here on the most important: I reject Moehler's argument for the weak principle of universalization, and hence the substance of his conclusions about the content of contractarian morality.

Moehler states his principle as follows:

---

[95] *Ibid*.: 17.
[96] *Ibid*.: 17-18, 114-129.
[97] Again, with the obvious exception that his concern is moral contractarianism and mine is political contractarianism.
[98] On each of these points, see the relevant remarks in the Introduction.
[99] See the discussion in the Interface, and in Chapters 5, 6, and 7.
[100] In the text, I discuss differences flowing from the distinction between moral and political contractarianism, the question of inclusion, the baseline for bargaining, the role of coalitions, and the basis of equality in contractarian theories. I note here several less significant ways in which my account diverges from Moehler's. At the metaethical level, Moehler (2018: 21, 26) appears to follow Raz. I argue against this view in §2.4.1 (*infra*). We also differ on the role of actual versus hypothetical agreement in social contract theories (Moehler, 2018: 11, 33, 178, 217; compare §1.2, *supra*, and the discussion in the Interface, *infra*). Moehler suggests that peace can sometimes take moral priority over justice (2018: 17-18, 101-102, 211-213), whereas I reject this claim (see the Introduction, *supra*). We differ in our characterisation of unreasonable agents (Moehler, 2018: 18, 97, 101-102, 104; compare §2.4.3, §3.1 n19, §3.2.1, and §5.5, *infra*). We disagree about the justifiability of imposing one's own moral views on fellow citizens (Moehler, 2018: 26-27; compare the remarks in the Introduction, *supra*). We diverge somewhat on the degree to, and manner in, which contractarian principles have universal scope (Moehler, 2018: 27, 164; compare §4.1 and §7.1, *infra*). Moehler endorses a version of the Nash bargaining solution (2018: 54-56, 84-88, 127-129), whereas I am sceptical of it (§6.5, *infra*). We differ in our handling of tuistic preferences (Moehler, 2018: 53, 102; compare §5.1, particularly n4, and the discussion of covariance in §7.6, *infra*). We also differ somewhat in our willingness to be revisionary about the standard commitments of rational choice theory (Moehler, 2018: 17, 23, 57-58, 104-105; compare §8.4, *infra*). I will comment further on these differences as and when they arise.

> "In cases of conflict, only pursue your interests subject to the constraints that your opponents can (i) enter the process of conflict resolution at least from their minimum standards of living, if the goods that are in dispute permit it, and (ii) fulfill their interests above this level according to their relative bargaining power."[101]

This requires some unpacking, but the key issue is with constraint (i), with whether or not it is redundant, and if it is not redundant, with whether or not it can be defended on contractarian grounds.

Moehler argues that rational agents in his hypothetical choice situation would insist that any candidate principle for conflict resolution "grants [each of] the agents at least the means that they need to maintain their existence as separate agents and satisfy their basic human needs…"[102] He summarises this as a requirement that the principle selected guarantee 'minimum standards of living'. These are to include "certain basic rights that protect equally the agents' physical integrity and personhood and, in this sense, their existence as separate autonomous beings... so that the interests of no member of society can be excluded or generally dominated by others…"[103] In the context of "capitalist democratic societies… basic rights may also include rights to private property, due process, equality of opportunity, and democratic participation…"[104] The guarantee of minimum standards of living will also include provision of any necessary material resources.[105] Later, he adds to these conditions that it is not sufficient for a society to *maintain* an agent at the threshold level represented by minimum standards of living; rather the second constraint of the weak principle of universalization requires that minimum standards of living are treated as a *baseline from which* agents must *gain* in proportion to their bargaining

[101] 2018: 125.
[102] *Ibid*.: 115. He assumes here a condition of moderate scarcity, such that satisfying the condition is possible within the resource constraints under which the agents are operating.
[103] *Ibid*.: 115-116.
[104] *Ibid*.: 116.
[105] *Ibid*.: 116-117. Moehler brackets discussion of people with disabilities who may need additional resources to meet their specific basic needs.

power.[106]

All of this sounds like a radical constraint on the contractarian theory. (Universal civil and political rights! Democracy! An unconditional basic income!) While Moehler's constraint might not go quite as far as some liberal egalitarian accounts, it appears to go further than many existing constitutional regimes, and it claims to do so without relying on any moral premises.[107] These considerations should give us pause: Perhaps I am reading too much into Moehler's remarks and the thresholds he has in mind are much more minimal than I have suggested (despite his somewhat expansive language).

Moehler appears to intend minimum standards of living to represent a non-agreement point for his theory. He writes, "The basic social, economic, and political goods that must be provided to all members of society… define, under the circumstances described, the disagreement point for the domain of pure instrumental morality."[108] I will later argue – *contra* Gauthier – that the appropriate non-agreement baseline for contractarian theories is Hobbesian rather than Lockean.[109] What this will amount to is the claim that contractarian bargaining theories must take as their non-agreement point the state of affairs in which all agents maximally employ their individual capacities in pursuit of their ends (including their capacities to physically force other agents to do things, but excluding any threats, and excluding any coordination between agents), and must guarantee all (included)

[106] *Ibid*.: 119-120. Moehler seems to assume that once one *has* minimum standards of living, one's relative bargaining power cannot be zero, but his reasoning on this point is not clear.
[107] Despite the fact that I defend substantively liberal egalitarian political institutions later, Moehler's constraint goes further than even I do in respect of the apparently persistent equal rights he describes. See §7.8, and particularly n74 and n75.
[108] 2018: 117-118; compare *ibid*.: 51. Where Moehler refers to a 'disagreement point' I refer to a 'non-agreement point'.
[109] §7.5. Compare Vanderschraaf, 2019: 276, 311. The commitment to the Lockean baseline is one of the points at which Moehler argues that Gauthier allows illicit moral considerations to enter his argument (Moehler, 2018: 51-52; see also Vanderschraaf, 2019: 310-311, 320). While I will argue that Gauthier is mistaken to adopt the Lockean baseline, and that the Lockean baseline can only be defended on moral grounds, I think Moehler mischaracterises Gauthier's argument when he suggests that it has a moral basis. Gauthier's appeal to the Lockean baseline relies on his account of constrained maximisation, which he takes to be grounded purely in (enlightened) prudential reason. Gauthier's argument can be mistaken without making the mistake of moralising.

agents at least this 'security' level.[110] I argue that, for most agents, this level will probably be very low indeed – a Hobbesian baseline, after all, is a life that is "solitary, poore, nasty, brutish, and short."[111] While I have not yet made this argument, I take it the appeal for the contractarian of this characterisation of the non-agreement point is fairly obvious: It assumes no moral constraint whatsoever on the use of force by agents to advance their ends. Given Moehler's description of minimum standards of living, he must think either that life in the Hobbesian natural state is much nicer than generally assumed, such that it guarantees rights and goods of the kinds he lists, or he must think that the non-agreement point is not violent conflict. The former is not plausible, given how Moehler defines cases of conflict – the whole point is that agents are willing to kill to get their way – so what could motivate the latter?

What seems to be causing the problem here is that Moehler confuses the idea of a non-agreement (or disagreement) point with the idea of a minimally acceptable settlement point. In a number of places, he argues that it would not be rational for agents to agree to a principle that would give them less than minimum standards of living.[112] But a non-agreement point and a minimally acceptable settlement point are quite different things. The latter is a (partial) specification of the distribution of the bargaining surplus. The former describes the absence of that surplus. Put in these terms, the explanation for the difference between my characterisation of the non-agreement point and Moehler's description of minimum standards of living should be clear. Nothing Moehler says motivates minimum standards of living *as a non-agreement point*.

Now, however, a difficulty arises for Moehler's argument. If minimum standards of living are not a characterisation of the non-agreement point, but are rather a constraint on the distribution of the bargaining surplus, then they may conflict with the *other* constraint Moehler places on the distribution of the bargaining surplus:

[110] §6.1, §8.5.
[111] Chapter 13 of *Leviathan* (1996: 89).
[112] 2018: 115, 116, 118, 119-120.

distribution in proportion to bargaining power. Either distribution in proportion to bargaining power will necessarily satisfy the minimum standards of living constraint, in which case the latter is redundant. Or, distribution in proportion to bargaining power will not necessarily satisfy the minimum standards of living constraint. In the latter event, which constraint prevails? If the proportionality constraint prevails, the minimum standards constraint is again redundant. If the minimum standards constraint prevails, *on what non-moralised, contractarian-consistent basis does it do so?* To put this another way, what Moehler appears to have done with the minimum standards of living constraint is reinvent the Lockean proviso he criticised in Gauthier – a point which is treated *as if* it were a non-agreement point for the purpose of allocating the remainder of a surplus, while not actually being a non-agreement point.

Can Moehler give a non-moralised defence of a non-redundant minimum standards of living constraint? As far as I can see, any attempt to do so will rest on two fatal equivocations. First, Moehler insists that agents are aiming to secure peaceful, long-term, stable cooperation.[113] This sounds lovely. But note that, if we're talking about practically stable social arrangements, over reasonably long time periods, which do not involve high levels of violent conflict, these conditions can be perfectly well satisfied by viciously oppressive regimes, including those incorporating slavery, feudalism, forced labour, and apartheid.[114] The idea that such regimes can be excluded simply by a requirement of mutual advantage, relative to a non-moralised baseline, is wishful thinking. But, if I am right about that, the aim of peaceful, long-term, stable cooperation is insufficient in itself to support the claims Moehler makes for minimum standards of living.

There is one more possible response. If we assume that our agents are roughly equal in their powers, there is a non-moralised route to the robust minimum standards of

[113] *Ibid*.: 114, 115, 116, 117-118, 119-121. On stability in particular, see *ibid*.: 120-121.
[114] As Rousseau notes, dungeons are peaceful (*The Social Contract*, Book I, Chapter 4 (1968: 54)). I have been at pains to distinguish 'cooperation' and 'coordination' for just this reason (see §1.5, and particularly n37, and the discussion of *modus vivendi* and moral minimalism in the Introduction, particularly n21, *supra*). Compare Vanderschraaf, 2019: 305-307. On the distinction between 'practical' stability and 'normative' stability (stability for the right reasons), see §3.1, *infra*.

living Moehler describes (equal rights and opportunities, political equality and democracy, guarantees of material wellbeing). Equally powerful agents, even bargaining from a non-moralised baseline, will agree to political institutions which leave them more or less equal.[115] As Moehler says, if we assume "a state of natural equality among agents, rational agents would not give up the defense of their interests by violent conflict resolution to ensure peace without receiving in return… basic rights that protect their autonomy to express their interests equally… so that the interests of no member of society can be excluded or generally dominated by others…"[116] This would still make the minimum standards of living constraint redundant (what's doing the work is the equality assumption), but perhaps we can read Moehler as using the minimum standards constraint simply to pick out one important implication of the more generally applicable proportionality constraint under conditions of natural equality.

The second equivocation, then, is more subtle, and goes to the heart of the difference between my account and Moehler's. What reason do we have to suppose that our agents are equally powerful? We can't stipulate this for moral reasons, as a contractualist would. Rather, Moehler follows Hobbes in appealing to 'natural' equality.[117] As Hobbes puts it:

> "Nature hath made men so equall… the difference between man, and man, is not so considerable, as that one man can thereupon claim to himselfe any benefit, to which another may not pretend, as well as he. For as to the strength of body, the weakest has strength enough to kill the strongest, either by secret machination, *or by confederacy with others*, that are in the same danger with himselfe."[118]

For Hobbes, the natural equality which makes people equally vulnerable in the pre-political state is contingent, in part, on the ability of weaker agents to combine with

[115] I discuss this in much more detail in §7.8.
[116] 2018: 115-116. See similar remarks at *ibid*.: 120.
[117] *Ibid*.: 8, 13, 114-115, 116 n37, 138.
[118] Chapter 13 of *Leviathan* (1996: 86-87, emphasis added). See also §7.5 n29, *infra*.

one another – to form *coalitions* – against the stronger. Absent this feature of the Hobbesian state of nature, natural equality is not a plausible empirical assumption. Unlike some, Moehler is aware of this. He doesn't truncate Hobbes's claim to neglect the confederacy clause.[119] Nonetheless, he excludes consideration of coalitions, and of cases with more than two agents generally, from his analysis.[120] I contend that, in doing so, he deprives himself of any non-moralised basis of equality for his theory. Without a basis of equality, he cannot defend the minimum standards of living requirement in a manner consistent with his contractarian methodology.

For Moehler's *moral* contractarianism, the absence of a non-moralised basis of equality will result in the failure of his argument, because without a basis of equality he will not be able to generate the impartiality required by genuinely moral principles. Read as a *political* theorist, Moehler must either accept moral minimalism, and hence give up his contractarianism; or, if he remains a political *contractarian*, he must provide an alternative basis of equality or significantly revise the substance of his conclusions about the content of contractarianism. One of my principal aims in Part II of this thesis is to defend a genuinely non-moralised, alternative contractarian basis of equality.

§1.6.3 Vanderschraaf's conventionalist contractarianism

The second recent account I want to examine is by Peter Vanderschraaf.[121] Vanderschraaf is categorised by Moehler as a moral conventionalist, rather than a moral (or political) contractarian, but this is something of a simplification.[122] Vanderschraaf is concerned with the interplay between conventions and more obviously contractarian elements such as bargaining problems.[123]

---

[119] 2018: 12, 42, 97, 108.
[120] *Ibid*.: 108-109.
[121] 2019.
[122] 2018: 12.
[123] For example, 2019: 149-175. Particularly striking is Vanderschraaf's endorsement of what he calls 'Baseline Consistency', since it seems to commit him to a kind of path-independence that is not at all typical of conventionalist approaches (*ibid*.: 311-318).

Vanderschraaf expressly sets out to defend a conception of justice as mutual advantage.[124] He argues that 'justice-as-convention' is "the most cogent characterization of the general theory of justice as conceived of as a system of rules for mutual benefit, or *justice as mutual advantage*…"[125] Vanderschraaf's understanding of a 'convention' is sufficiently broad to be accepted by contractarians. A convention, for Vanderschraaf, is "a system of strategies that characterize an *equilibrium* solution to a problem of coordination that has a plurality of such equilibrium solutions."[126] As with all of the other accounts I consider, and with my own approach, Vanderschraaf is especially concerned with conventions applied to problems of 'conflictual coordination' – cases in which there is the possibility for both conflict and mutually beneficial coordination. These conventions, on his account, are conventions of justice.[127]

Vanderschraaf begins with two questions he takes from Plato: What is justice? And, why should one be just?[128] Unlike many contemporary theorists – mainstream contractualists, for example – Vanderschraaf takes these questions to be deeply entangled. One of the main strengths of a justice as mutual advantage theory is that, given a mutual advantage answer to the first question, the second question becomes much more tractable. If justice is to one's advantage, it is not so mysterious why one has reason to be just.[129]

---

[124] *Ibid*.: xi, 323.
[125] *Ibid*.: xi, emphasis in original. Thus, on Vanderschraaf's account, justice-as-convention falls within the larger set of justice as mutual advantage theories.
[126] *Ibid*.: xiii, emphasis in original. See also *ibid*.: 81-81, 325-332. It remains an open question quite how determinate one might want a contractarian theory to be (see the discussion of Gauthier's political contractarianism in §1.6.1, *supra*). Contractarians in the narrower sense of Moehler (2018: 12) tend to look for greater determinacy relative to typical conventionalists – one of the main tasks of Vanderschraaf's book is to argue that conventionalism, as he understands it, is not *so* broad as to be inadequate as a theory of justice (2019: xii, 304-319). So long as contractarians do not think that their account is fully determinate (and so long as they accept a stability requirement, understood in terms of a game-theoretic equilibrium), they will strictly meet the criteria for a convention that Vanderschraaf has set out.
[127] *Ibid*.: xiii, 24-25, 44-84. Compare the Introduction, *supra*. For my discussion of the relationship between 'justice' (understood as a moral concept) and the object of contractarian agreement ('a set of normative principles to regulate political interaction'), see §1.7, *infra*.
[128] 2019: xi. Compare Moehler, 2018: 13.
[129] Vanderschraaf, 2019: xii-xiii, 319-321.

It will be helpful to pick out two different aspects of Vanderschraaf's second question, one general and one more particular. Why, we might ask, do we *in general* have reason to be just? To put this another way, what class or category of reason can we have to be just? It is to this question that mutual advantage theorists answer with 'prudential reasons'. However, we might also ask, why *in any particular case* do we have reason to be just? That is, what explains the stringency of our reason to be just, when in some cases it appears that we can gain more advantage by departing from justice? This particular aspect of Vanderschraaf's question identifies a potential problem with appealing to prudential reasons as the grounds for being just: it seems improbable to suppose that it will always be the case that our ends align happily with the requirements of justice. This is the compliance problem.[130] Much of Vanderschraaf's discussion is focused on the compliance aspect of his second question.[131] The point of requiring a convention to be a (game-theoretic) equilibrium is that, by definition, agents will have no incentive to deviate (unilaterally) from an equilibrium. The convention will have an important kind of stability. Equilibrium solutions solve the compliance problem (at least so long as one assumes agents are acting independently, rather than agreeing to coordinate in some way).[132]

Social contract theories, for Vanderschraaf, are studies of resource division (or 'bargaining') problems.[133] In such problems, agents make claims for some share of a resource which is produced by the coordination of the agents. The set of claims made can either be compatible (the sum of the claims is not greater than the total of the available resource) or incompatible (the sum of the claims exceeds the total of the available resource). If the claims are incompatible, agents get nothing; the

[130] See the remarks in the Introduction, and particularly n35.
[131] 2019: 18-25, 67-77, 127-159, 175-189, 193-215, 226-270, 287-292.
[132] The concept of an equilibrium emerges from the development of noncooperative game theory (*ibid*.: 18-25). In noncooperative games, agents lack commitment power; among other things, they lack the capacity to make binding agreements (see Moehler, 2018: 103-107, 143-145, 172-176). It is generally thought that cooperative game theory (in which agents can make binding agreements) is not an appropriate tool for contractarian theories, because such theories are supposed to be grounding the (non-natural) institutions (such as the state, or the institution of promising) that empower agents to make binding agreements (*ibid*.: 108-109). It is supposed to follow, then, that if agents cannot make binding agreements, and are rational, they must act independently.
[133] 2019: 160ff. The following discussion assumes a basic knowledge of bargaining theory. I lay out the fundamentals of bargaining theory in Chapter 6, *infra*.

resource remains undivided and inaccessible. If the claims are compatible, the resource is divided among the agents according to their respective claims. Assuming the agents each want as much of the resource as possible, they will seek a Pareto efficient division, such that the sum of their claims is neither greater nor less than the total of the available resource. For a perfectly divisible resource, there are infinitely many efficient divisions, corresponding to all of the possible proportions in which the resource could be exhaustively allocated among the agents. The set of these divisions constitutes the Pareto frontier.

Given this structure for the problem, every point on the Pareto frontier is an equilibrium for the problem. Take a case of two agents, *L* and *M*, and a point on a frontier (80, 20) which allocates an 80% claim of a resource to *L* and a 20% claim of the resource to *M*. In this case, neither agent has an incentive to unilaterally alter his or her claim. Consider *M*. *M* has no incentive to claim less than 20% of the resource, because he wants as much of the resource as possible. He also has no incentive to claim more than 20% of the resource, *given* that *L* is claiming 80%. If *M* changes his claim from 20% to 30%, the sum of the claims of the agents will exceed the total of the available resource. The claims of *L* and *M* will then be incompatible, and *M* will get none of the resource (which is obviously less than 20% of the resource).[134] The situation of *L* is the same. And this is true of *any* point on the Pareto frontier. Which is to say, when considering a resource division problem with this kind of structure, requiring that a solution to the problem be a (game-theoretically) stable equilibrium radically underdetermines the solution. There are infinitely many equilibrium solutions to the problem.

[134] The problem, as presented, of course abstracts from a great many issues. As I have set it out, the problem assumes that moving away from an equilibrium on the Pareto frontier results in a payoff of zero. But a more complicated description of the problem might allow agents, when for some reason they do move away from an equilibrium, to make a new set of claims (which can then be either compatible or incompatible). Suppose the resource in question is a 'flow' rather than a 'bundle', such that the payoff of having the resource is a product of the duration over which it is held. It is then perfectly possible for it to be rational for *M* to collapse the initial (80, 20) equilibrium, if he can expect that the sum of the flow to him under a new equilibrium will be greater, perhaps over a longer duration (taking into account the opportunity costs of a flow of zero for some period between the collapse of the old equilibrium and the establishment of a new one), than the flow to him under the uninterrupted initial equilibrium. The simple problem presented in the text is thus purely illustrative.

This conclusion extends to theories of justice as mutual advantage generally. Given a plausible set of necessary conditions for characterising a theory as a theory of justice as mutual advantage – conditions which essentially mirror the conditions which define a bargaining problem – the theory will face an underdetermination problem of the kind described.[135] The central problem, then, for social contract theories is an equilibrium selection problem. Vanderschraaf considers two kinds of approaches to equilibrium selection: 'rational choice' approaches and 'salience-learning' approaches.[136] Rational choice approaches aim to add some further constraint to the set of conditions which characterise an equilibrium such that, given the constraint, all but one point on the Pareto frontier is excluded, and so the resource division problem has a *unique* solution.[137] They do this principally by appealing to one of three bargaining solutions: the Nash solution, the Kalai-Smorodinsky solution, and the (Kalai) 'egalitarian' solution.[138]

Rational choice approaches to the equilibrium selection problem argue for one of the three bargaining solutions over the others by one (but ideally both) of two means: 'axiomatic' and 'strategic' (also known as 'procedural').[139] Vanderschraaf rejects strategic/procedural arguments (which are usually considered to favour the Nash solution), because they assume too much common knowledge among the parties, and because they are often indeterminate for bargaining games involving more than two players or goods that are not perfectly divisible.[140] He rejects axiomatic arguments for the Nash solution and for the Kalai-Smorodinsky solution.[141] He also notes an objection to axiomatic arguments for the egalitarian solution, though he ultimately argues that this objection – rooted in the solution's dependence on the possibility of interpersonal comparisons of utility – is defeasible.[142]

---

[135] Vanderschraaf, 2019: 273-279, 304-308.
[136] *Ibid*.: 163-175, 308-311.
[137] For discussion of the concept of uniqueness in social contract theories, see the exchange between Thrasher (2014) and Thoma (2015).
[138] Vanderschraaf, 2019: 163-165. I give a full explanation of these solutions in §6.2 and §6.3, *infra*. These are not the only possible solutions, but they have been the most prominent in the literature.
[139] Vanderschraaf, 2019: 163. See §6.3, §6.4, and §6.5, *infra*.
[140] 2019: 166-167. See §6.4, *infra*.
[141] 2019: 165-166.
[142] *Ibid*.: 165, 185-187. See §6.2, *infra*.

Turning to salience-learning approaches to the equilibrium selection problem, Vanderschraaf deploys a series of computer simulations designed to model how real-world agents might go about developing a norm which can guide them when they encounter the problem selecting one equilibrium from many which are possible.[143] If, in a particular population, there is such a norm, it can provide a 'focal point' to aid agents otherwise facing an intractable coordination problem.[144] Vanderschraaf's simulations model a group of agents engaged in a series of rounds of interaction. In each round, each model agent is paired with another, and then must engage in a simulated resource division problem. The model agents are given initial rules for how to behave in the interactions, and also rules on how to adjust their claims in later rounds based on the outcomes they achieve in earlier rounds; that is, based on 'learning' about the claims other model agents in the population are making, and whether or not the rules a model agent is following are producing or are failing to produce compatible claims (and, hence, payoffs). The simulations contain a 'fitness' algorithm which reinforces the selection by the model agents of rules for making claims which increase the probability of making compatible claims. The simulations are then run for a large number of rounds, and we observe whether the models' 'inductive learners' converge on a stable rule for making claims. Having examined several different models, Vanderschraaf concludes that they tend to converge on equilibria at or near the equilibrium that would be identified by the egalitarian bargaining solution: "[W]hile many different bargaining conventions can emerge from trial-and-error learning, egalitarian solution equilibria may become psychologically prominent because these equilibria tend to be attracting points of such learning."[145]

Vanderschraaf argues that these results support a conclusion in favour of an egalitarian solution to the bargaining problem.[146] Nonetheless – and somewhat

[143] 2019: 167-175; see also *ibid*.: 137-145, 333-348.
[144] The seminal work on focal points as a response to the equilibrium selection problem in coordination games is by Schelling (1960).
[145] Vanderschraaf, 2019: 175.
[146] *Ibid*.: 186.

surprisingly, for someone identified as a conventionalist – he rejects a salience-learning approach as a *sufficient* argument for a solution to the problem of equilibrium selection.[147] While a salience-learning approach is good at *explaining how* norms of justice might arise, "it is simply not enough to say that justice as mutual advantage *is* a system of equilibria that are salient in a certain way or that are certain endpoints of learning."[148] The problem is that while, given the right initial conditions, a salience-learning approach might *tend* to produce equilibria which we would recognise as fair or cooperative, there is no guarantee that it will do so – it could also produce deeply unfair equilibria or could fail to produce any cooperative equilibrium.[149]

If a salience-learning approach is not sufficient, then at least some appeal will have to be made to a rational choice approach. Vanderschraaf briefly surveys some candidates, including theories developed by Rawls, Harsanyi, Scanlon, and Binmore, but he argues that all of these make use, to one degree or another, of a veil of ignorance device, which cannot be justified within the constraints of a justice as mutual advantage (*viz*., contractarian) methodology.[150] Gauthier's moral contractarianism is exposed to a similar objection, targeting his use of the Lockean proviso rather than a veil of ignorance.[151]

What Vanderschraaf needs is a constraint which he can add to the conditions defining the bargaining situation to generate a unique equilibrium, and which can be justified without appealing to a moral principle. To fill this role, he offers us 'Baseline Consistency'.[152] Suppose we stipulate some non-agreement point as a baseline and apply a bargaining solution to this point. Doing so selects a unique

[147] *Ibid*.: 308-309.
[148] *Ibid*.: 309, emphasis added.
[149] *Ibid*.: 308-309; see also Moehler, 2018: 124-125 n46. Note that a *political* contractarian version of Vanderschraaf's account would not make this objection to a salience-learning approach. The insufficiency, here, of a salience-learning approach is a result of Vanderschraaf's commitment to moral contractarianism.
[150] 2019: 309-310.
[151] *Ibid*.: 310-311. I have argued (§1.6.2, *supra*) that Moehler's minimum standards of living constraint is similarly flawed. On Gauthier's Lockean proviso, see §7.5, *infra*.
[152] 2019: 311-318. Compare Kalai, 1977.

point on the Pareto frontier of the solution space. Suppose, next, that there is some change in social circumstances, such that the bargaining surplus is altered in some way – perhaps there is more to distribute. Geometrically, this amounts to a change in the size (and perhaps the shape) of the solution space. There is now a question about how the gains are to be distributed. We have two options. We could take our present position (the point on the original Pareto frontier selected by our initial application of our bargaining solution) and treat it as a new non-agreement point, applying our bargaining solution from that point to select a unique point on the *new* Pareto frontier. Alternatively, we could return to our original non-agreement point and use our bargaining solution to recalculate a point on the new Pareto frontier, given the new solution space. A *baseline consistent* bargaining solution will select the *same* point on the new Pareto frontier, regardless of which of these options we take. Such a solution is path-independent – it doesn't matter whether we get to the new solution in one leap, straight from the original non-agreement point, or via any number of steps from one *status quo* point to the next whenever there is a change in bargaining surplus.[153]

Adopting baseline consistency changes how we should evaluate contending rational choice approaches to the equilibrium selection problem. The Nash bargaining solution is not baseline consistent, whereas the egalitarian bargaining solution is.[154] Vanderschraaf now has two independent arguments for accepting the egalitarian bargaining solution as an equilibrium selection device, one based on a rational choice approach and relying on baseline consistency, and another based on a salience-learning approach and supported by simulation. Nonetheless, even with the addition of baseline consistency, his criteria do not *reduce* justice as mutual advantage to the egalitarian bargaining solution.[155] However, if we accept his criteria, "the resulting set of equilibria always includes the set of egalitarian solutions

[153] This is precisely the kind of property for a bargaining solution that Moehler rejects (2018: 119, 120). In §7.5, I object to Gauthier's argument for the Lockean proviso in part because it is baseline inconsistent.
[154] Vanderschraaf, 2019: 312-318.
[155] Thus, Vanderschraaf's account is not fully determinate, and despite the role played by rational choice elements which are contractarian in Moehler's narrower sense of the term, it remains plausible to describe Vanderschraaf as a conventionalist.

and in fact reduces to the set of egalitarian solutions if the parties are assumed to have equal natural bargaining ability."[156] Vanderschraaf concludes that he has identified a theory of justice as mutual advantage which is stable (because it selects an equilibrium), which is sufficiently determinate (because baseline consistency narrows the range of eligible equilibria), which is consistent with contractarian methodology (because no moral elements are smuggled in), and which is genuinely just (because baseline consistency limits the range of eligible equilibria to those which are relatively egalitarian).[157]

As with the other accounts I consider here, there are many points at which I overlap with Vanderschraaf's account. For the purpose of this thesis, I have set aside the compliance problem, but Vanderschraaf's discussion of compliance questions is particularly sophisticated.[158] As before, I will focus on the points at which I diverge from his analysis. I will highlight two.

First, I reject Vanderschraaf's argument for the egalitarian bargaining solution. As I will argue in §6.2, given an assumption of deep pluralism, the shared understanding of utility levels which is necessary to implement the egalitarian bargaining solution is simply not available to the agents with whom I will be concerned, and is not consistent with a purely contractarian methodology.[159]

Second, Vanderschraaf is no closer to identifying a contractarian basis of equality than Moehler. Upon closer inspection, in fact, Vanderschraaf's conclusion is startlingly weak. Having noted that his solution criteria, including baseline consistency, *do not exclude* egalitarian solutions, he goes on to say that the set of eligible equilibria reduces to the set of egalitarian solutions *if* we assume agents have "equal natural bargaining ability".[160] It is not entirely clear what he means by "equal

[156] *Ibid*.: 318.
[157] *Ibid*.: 322-323.
[158] See, in particular, his analysis of the Indefinitely Repeated Provider-Recipient Game (*ibid*.: 280-292), which serves as his reply to the Vulnerability Objection (an aspect of what I have called the exclusion problem).
[159] Moehler makes the same objection (2018: 52-53, 85, 161).
[160] 2019: 318.

natural bargaining ability". His text, at this point, has a footnote in which he identifies the equal natural bargaining ability claim with the symmetry axiom (a common axiom in bargaining theory). This suggests Vanderschraaf is drawing on Nash's original use of the idea of equal natural bargaining ability.[161] But Nash later clarified that symmetry has nothing to do with bargaining ability (as was in fact clear from the formal components of Nash's original paper – the confusion arose in his informal discussion).[162] Alternatively, Vanderschraaf might be using "natural bargaining ability" to refer instead to bargaining *power*.[163]

If, in this passage, Vanderschraaf is referring to symmetry, then nothing in his claim rules out agents having unequal bargaining power (symmetry is compatible with inequalities in bargaining power). If agents can have significantly different levels of bargaining power, there is no reason to think that Vanderschraaf's criteria will restrict the eligible equilibria to those which are relatively egalitarian.[164] If, on the other hand, Vanderschraaf is referring to equal bargaining power, then he will beg the question. His claim will be that, *if* we assume agents have equal bargaining power, justice as mutual advantage returns relatively egalitarian results. But he will have given us no reason to assume agents have equal bargaining power, and there is no obvious non-moralised reason to grant the assumption. Vanderschraaf will end up in exactly the position he has put Rawls, Harsanyi, Scanlon, Binmore, and Gauthier – introducing moralised premises which are inconsistent with a purely contractarian methodology. As I have indicated, one of the principal tasks of my argument will be to defend a non-moralised basis of equality which can fill this gap in contractarian

[161] Nash, 1950: 159. For a specification of the symmetry axiom, see §6.4, *infra*. In fact, Vanderschraaf's footnote (2019: 318 n75) says that he is *not* assuming equal natural bargaining ability, so it is not clear what the point of the remark in the main text is. The note could be read as implying that he is assuming *relatively limited differences* in bargaining ability, rather than no differences, but no explanation of, or non-moralised justification for, such an assumption is given.
[162] Nash, 1953: 137-138. The text of Vanderschraaf's footnote (2019: 318 n75) echoes Nash's phrasing in this later paper (1953), which identifies symmetry with something like an anonymity condition (compare Thrasher, 2014: 693; see §8.3.1 n34, *infra*). But it is then unclear why Vanderschraaf persists in linking symmetry to bargaining ability (as he also does at 2019: 220 n75).
[163] Vanderschraaf earlier appears to conflate symmetry and equal relative power (2019: 221 n76). For my discussion of bargaining power, see §6.3, the introductory remarks in Chapter 8, and §8.3.1, *infra*.
[164] Perhaps Vanderschraaf could appeal to Hobbes's 'rough equality' assumption (see the discussion in §1.6.2). His argument for the salience of democracy as a focal point equilibrium relies on this assumption (2019: 221). But, in that case, the response is the same as the one I gave when responding to a similar move by Moehler (§1.6.2, *supra*).

arguments.

§1.6.4 Muldoon's experimental contractarianism

The final example I will consider of a recent contractarian theory is presented by Ryan Muldoon.[165] Even this opening sentence may be controversial. On first inspection, it is not at all clear that Muldoon can sensibly be described as a contractarian, whether moral or political. Central to the motivation of Muldoon's project is a Millian concern with *progress* – with the iterated refinement over time of a person's (or a society's) understanding of justice and the good life.[166] Muldoon's question is less "What is justice?" and more "How can we best go about organising society so as to improve our understanding of justice?" As much as anything else, Muldoon appears to be setting out a (social) moral epistemology, rather than a political theory. He follows Mill in thinking that citizens undertaking different experiments in living will generate useful data, which other citizens – individually and collectively – can then use to inform their moral judgements. Justice, for Muldoon, is not an end point or ideal state, it is a *trajectory*.[167] He argues for a political philosophy oriented to *discovery* rather than *justification*.[168] Mainstream political philosophy – most prominently Rawlsian contractualism – focuses on justification, attempting to identify and defend ideal, universal, and timeless principles to govern social and political life.[169] Discovery, on the other hand is a never-ending process by which we repeatedly reassess our social and political institutions in light of new evidence, and then adjust our institutions in response to this evidence.[170] Commitments of these kinds look flatly inconsistent with a contractarian methodology, not least in their apparent perfectionism.

---

[165] 2016.
[166] *Ibid*.: Introduction, Chapter 1, §2.3, §2.8, §3.4, §5.3.1, Chapter 6; compare Mill, *On Liberty* (1974).
[167] 2016: Introduction §5.
[168] *Ibid*.: §2.4, §2.8, §3.4, Chapter 6.
[169] *Ibid*.: Chapter 1, §2.3, §3.2, Chapter 6. See the discussion of universalism in Chapter 4, *infra*.
[170] 2016.: Introduction, Chapter 1, Chapter 4, Chapter 6. Muldoon (*ibid*.: Chapter 1) explicitly endorses Sen's scepticism of (his understanding of) ideal theory (2009).

This initial picture is incomplete, however. The animating concern, in Muldoon's account, is with *diversity*.[171] No less than Moehler, Muldoon is concerned with diversity in the sense of deep pluralism about values.[172] More unusually, he is also concerned with diversity of *perspectives*.[173] As he explains:

> "Perspectives are simply the filters we use to view the world. This supposes that we do not take in the world "as it is," but instead (consciously or unconsciously) choose to group certain features together, choose to ignore certain information while focusing on other information, and choose systems of representation and interpretation. Perspectives are thus mental schemata that embody those choices."[174]

Perspectives have two components: An *ontology* which determines how one distinguishes and groups objects and states of affairs. And a set of *heuristics* which determine how one evaluates the elements identified by the ontology.[175] Muldoon's insight is that differences in perspective – particularly those driven by different ontologies – can generate disagreement between people even if they seem to agree about values.[176] This possibility adds an additional source of pluralism with which social contract theories must contend.

In aiming to develop an account of the social contract which can address this kind of deep diversity, Muldoon is working squarely within the theoretical space of traditional contractarians.[177] Moreover, as with traditional contractarians, Muldoon's response to the problem posed by deep pluralism is to search for a normatively minimal theory – a theory citizens can use to give one another reasons, while

---

[171] 2016: Introduction.
[172] §1.6.2.
[173] 2016: Introduction §2, §2.2, §3.4.2, §4.3.
[174] *Ibid*.: §3.4.2.
[175] *Ibid*.: §3.4.3.
[176] For example (*ibid*.: §4.3), we might all agree that murder is wrong, but disagree about what counts as a murder. Alternatively, consider the categories 'domestic violence' (*ibid*.) or 'sexual harassment'. Muldoon argues that agents who *have* these categories in their ontologies are unlikely to evaluate certain actions in the same way as agents who *lack* the categories. I discuss a related problem in §2.3.3, *infra*.
[177] 2016: Chapter 6.

nonetheless having extremely diverse perspectives and values. Like many traditional contractarians, Muldoon appeals to bargaining theory to fill this role, and for exactly the same reason: bargaining requires minimal (though not zero) ontological and evaluative agreement among agents.[178] Read in this light, Muldoon's project is less a Millian quest for increasingly precise approximations of perfect moral knowledge, and more a matter of asking, "Given that circumstances may have changed since last we asked, what is justice for us, in our present circumstances, for the time being?" The key components of Muldoon's project are a context of deep pluralism, the use of a normatively minimal bargaining theory to navigate said pluralism, and a commitment to periodic revision (*dynamism*). These are classic contractarian commitments.[179]

It might be thought that a commitment to periodic revision of the social contract is at the core of Muldoon's account but is *not* a feature of contractarian accounts – that contractarian accounts follow contractualists in seeking final, *static* answers to questions of political organisation.[180] This is a mistake. Contractarian accounts are *modus vivendi* theories.[181] Any *modus vivendi* theory is necessarily committed to adjusting the political institutions it prescribes in light of changing circumstances.[182] In fact, as we'll see, one of the main objections to such theories is that they are problematically unstable – *too* responsive to changing circumstances.[183] Of course, contractarian theories do aim to identify principles which *constrain* the set of justified political institutions.[184] But so does Muldoon.[185] There may be differences of degree here, but not of kind.

---

[178] *Ibid*.: §4.3. See, in particular, Muldoon's Chinese market and prison economy (*ibid*.: §4.3 n1) examples.
[179] It is not clear to me, and I do not mean to suggest, that Muldoon would reject (what I have called) the Millian reading of his project. For my purpose, it is sufficient that the Millian motivation is detachable from the overtly contractarian elements of Muldoon's account. In this sense, his account is political, in that it is consistent with, but does not require, the moral realist underpinnings provided by the Millian motivation.
[180] Muldoon seems to think as much, for instance (2016: Introduction §3, Chapter 1, §2.8, Chapter 6).
[181] See the Introduction, §3.2.2; on dynamism in a contractarian context, see §7.8.
[182] The same concern is implied by Vanderschraaf's discussion of decomposability (2019: 311-318).
[183] §3.2.2.
[184] See §1.6.3, n123.
[185] 2016: §5.3.2, §5.4.

Reading Muldoon as a contractarian, there are three aspects of his approach to bargaining theory it will be helpful to examine in greater detail. The first is his stance on idealisation. Muldoon rejects any form of idealisation that would eliminate the diversity of perspectives among agents which is his animating concern. This is one of his main objections to contractualist theories.[186] He argues that one consequence of agents in his bargaining situation having different perspectives is that they will have preference orderings over alternative sets of social arrangements which are *incomplete*.[187] This is potentially a problem because standard bargaining theory (including the Nash solution endorsed by Muldoon) requires agents to have complete preferences.[188] As it happens, I think Muldoon exaggerates the worry about incompleteness.[189] But when faced with a choice between maintaining a diversity of perspectives which accurately represents agents in the world as they actually are, and idealising agents in such a way as to satisfy the requirements of rational choice, Muldoon picks the former. Speaking about his use of a bargaining solution more generally, he says:

> "I should… note that I don't mean for the Nash Solution to do much *normative* work in my account. There are a number of alternative bargaining solutions, and while… I think that the Nash Solution has a number of features that make for an appealing bargaining solution, largely I am using it as a descriptive claim about where agents *will* end up in a bargain, rather than a normative claim about where they *should* end up. The normative significance here comes from the fact of agreement amongst the parties in the bargain, not that the agreement

[186] *Ibid*.: §3.2.
[187] *Ibid*.: §4.5.1. For a discussion of the concept of a preference, and of completeness, see §5.1, §5.2, and §5.3, *infra*.
[188] 2016: §4.5. Note that while Muldoon endorses the Nash solution, he is open to alternatives (*ibid*.).
[189] The examples he gives (*ibid*.: §4.5.1) look like they will result in agents being indifferent over alternative options for which, due to their perspective, they cannot distinguish all potentially relevant elements. In his 'dragonfruit' example, he denies that he can *feel* indifferent with respect to a fruit of which he has never heard. But this looks like an equivocation over the formal understanding of 'indifferent' and the colloquial understanding (see §5.3, *infra*). The presence of (formal) indifference relations in an agent's preference set does not imply the agent has incomplete preferences. If Muldoon anticipates other routes to (formal) incompleteness, he does not articulate them.

is characterized by the Nash Solution rather than some other solution."[190]

This is a puzzling passage. Leaving aside the ambiguity over exactly how much work 'not much work' entails, why is the *fact* of agreement supposed to have *normative* significance? I can see why, *if* one is committed to certain moral premises about respect for autonomy and keeping promises, one might then think that the fact of agreement is normatively significant. But such commitments are incompatible with a contractarian method. Alternatively, perhaps Muldoon's account doesn't much care about the content of an agreement, but only that there *is* agreement, and hence social stability (which itself is valued by all agents, let's suppose). If so, we seem to have given up entirely on the idea that the social contract can be a device of justification. If the mere fact of agreement is sufficient, the constraints Muldoon *does* endorse for his bargain (some combination of Pareto efficiency, proportional gain, and equalised burden sharing for externalities produced by provision of public goods) look redundant.[191] If we're not idealising to *some* extent, why even insist that a bargain be Pareto efficient? If we are going to idealise to at least some extent, I will argue that we should require (among other things) that agents have complete preferences, and more generally that they meet a set of conditions characterising rational choice. On my view, the normative significance of (hypothetical) agreement is that it tracks the reasons actual agents have.[192]

The second aspect of Muldoon's approach to bargaining theory which is of note is the attention he pays to coalitions. Given his concern with diversity, it is not surprising that Muldoon is interested in the practical limits that diversity might place on a social contract theory. He argues that, in diverse societies, citizens will generally fail to see *all* others as moral equals (at least absent the moral education provided by living under the institutions he recommends).[193] Rather, the society will divide into various in-groups and out-groups. Under these circumstances, a problem

[190] 2016: §4.5, emphasis in original. Vanderschraaf (2019: 169-175) would question the descriptive credentials of the Nash solution to which Muldoon appeals here.
[191] 2016: §5.3.2, §5.4.
[192] For further discussion of idealisation, see the Interface. For discussion of hypothetical agreement, see §1.2.
[193] 2016: §5.1.

of 'stability' arises for contractualist theories like Rawls's. While, under the veil of ignorance, citizens might agree to distributive principles which require extensive public welfare programmes, once the veil lifts and they see that the less advantaged in society are members of various out-groups (ethnic and religious minorities, immigrants, etc.), citizens will not actually be motivated to implement Rawlsian principles of justice.[194] The upshot for Muldoon is that, if we want our theories to give real citizens motivationally-effective reasons, then we need some way of deciding (i) who gets represented in a bargaining situation and (ii) the extent to which they get represented equally, which will be acceptable to actual agents with their various diverse perspectives (rather than to hypothetical agents, behind a veil of some kind).

The solution Muldoon proposes is a conception of *minimal equality*.[195] In standard contractarian fashion, he begins by returning to Hobbes's rough equality claim.[196] However, unlike many others, Muldoon emphasises the point that Hobbes's claim is based on all people being vulnerable at least to *coalitions* of other agents, and not simply to single other agents.[197] Hobbes's claim is descriptive, and so will be admissible by agents from diverse normative perspectives. Muldoon extends Hobbes's analysis of vulnerability in two ways. First, he notes that the claim about vulnerability applies to groups of agents as well as to individuals. He argues that families, communities, and religious groups will each themselves be vulnerable to some more powerful coalition.[198] Second, he applies the Hobbesian analysis to economic productivity as well as to a capacity for violent attack. That is, he argues that for any particular individual or group, there will be some coalition of other

[194] See the discussion of the problem of alienation in the Interface, and of the idea of representation in Thrasher (2019). Muldoon's analysis at this juncture seems to rest on a mistaken 'external' understanding of Rawlsian liberalism, but that doesn't undermine the more general point he is making. On the internal/external distinction in analyses of Rawlsian liberalism, see Quong (2011: 138-144). I discuss stability in some detail in Chapter 3, *infra*.
[195] 2016: §5.2.
[196] See §1.6.2 and §1.6.3 (particularly n164).
[197] 2016: §5.2 (particularly n3).
[198] There is a fairly obvious objection threatening at this point. If the group in question is some particularly dominant ethnic group, and members of the dominant group vastly outnumber all other ethnic groups combined, the Hobbesian vulnerability claim (at the group level) will fail. Thus, there is a limit on the scope Muldoon has to extend the Hobbesian analysis from individuals to groups.

agents with equal or greater productive capacity:

> "What we find is that every individual is equal insofar as her economic bargaining power can be equaled by some competing coalition. It is important to recognize two features of this. One, that more productive agents create more wealth, and as such have a superior bargaining position relative to less-productive agents. But two, these more productive agents have a lot more to lose relative to the median agent if they are unable to find a bargaining partner, so relative to the set of possible partners acting as a coalition, they are in a weaker position, especially in the long run."[199]

Parsing these remarks is not straightforward, but Muldoon appears to be saying a couple of things. First, it seems plausible to suppose that agents with greater economic productivity will (*pro tanto*) have more bargaining power when negotiating distributive claims. However, second, if we assume that even very productive agents require the assistance of others to realise certain gains, the productive agents have a lot to lose if they fail to find the assistance they need (they have more to lose – or, more precisely, forego greater gains – than the people with whom they are bargaining). Muldoon's claim here seems to be about the greater opportunity cost for productive agents of failing to find a bargaining partner.[200] This, he says, tends to offset the greater bargaining power of the more productive agents. He gives, as an example, the balance of bargaining power between trade unions and employers. While individual workers may have little bargaining power relative to employers, employers are totally reliant on the coalition of workers for production, and so the coalition (the union) has a great deal of bargaining power. This (relatively) equal bargaining power depends on the solidarity of the workers in the union, which is why solidarity is so important.[201]

---

[199] *Ibid*.: §5.2. The reference to the 'long run' refers to the distinction between one-shot games and repeated games. Muldoon claims that it is harder to use "perceived bargaining advantages" to one's own benefit in repeated games, but his analysis on this point isn't very clear.
[200] If this is right, he seems to be making the kind of argument Gauthier (1986: 134-139) does in appealing to Zeuthen's account of rational concession. See §6.4, *infra*.
[201] 2016: §5.2, particularly n4.

I think Muldoon has an important insight here, but there are several problems with his argument. First, it looks like it fails on its own terms. The problem with which he was concerned was a problem of political exclusion generated by the tendency of people to separate into in-groups and out-groups. What his analysis shows is that even the most productive agents have reasons to want to enter a social contract, and that there are limits to the scope of inequality among those included in a social contract: "No individual can [sensibly] opt out of the bargain, and no individual can dominate the bargain."[202] He has not shown that productive agents or coalitions have reasons to include less productive agents in the social contract. Moreover, given this failure to ground inclusion, and given the claim that life outside the social contract is desperately poor, the degree of inequality licensed by the social contract will remain high, because less productive agents threatened with exclusion will agree to almost anything to be allowed inside.[203]

The second problem with Muldoon's argument is the role being played by solidarity. He is clearly aware that his coalition-based argument for even minimal equality depends on treating the coalitions of more vulnerable agents as more or less monolithic. But such coalitions are made up of individual agents who can be tempted to defect from the coalition in return for a share of the social product – strikebreakers, scabs, and blacklegs, to follow Muldoon's labour union analogy. A contractarian certainly can't simply assume solidarity. Perhaps it could be argued that defectors would be undermining their own interests in the long run. But for this kind of reply to go through, it seems Muldoon would have to help himself either to a conception of collective rather than individual rationality, or to the assumption of some kind of commitment power.[204] Either would be problematic in a contractarian theory.

The third aspect of Muldoon's approach to bargaining theory I want to examine is

[202] *Ibid*.: §5.2.
[203] *Ibid*. Compare the discussion in §7.5, *infra*.
[204] On the distinction between individual and collective rationality, see Vallier (2014: 148). See Moehler (2018: 103-106, 172-174) on the use of commitment power in contractarian models. Gauthier gives an example of coalition dynamics involving a kind of commitment power (1988: 396-397). I discuss his proposal in §8.2.2, *infra*.

his account of the baseline. As I have noted, one of the constraints Muldoon endorses on his bargaining solution is a principle of proportional gain. More precisely, he says that:

> "[G]ains from social cooperation should be distributed in proportion to contributions to production. This principle is based on two ideas. First, that insofar as we are treating society as a system of social cooperation, we must ensure that everyone engaged in that cooperation benefits from the fruits of that effort. The second idea is that *no group should be required to subsidize another group to the extent that they would have been better off on their own*."[205]

Muldoon gives us a variety of glosses on this idea, but his discussion of it contains an important ambiguity that he does not successfully resolve. The proportional gain principle necessarily relies on a conception of a baseline with respect to which contributions and gains are measured. The baseline Muldoon proposes is counterfactual.[206] One's contribution to a social product is the difference between the social product realised with one's contribution and the social product which would have been realised by other agents *if one had been absent*. The problem with taking the *absence* of an agent as the relevant counterfactual is that an absent agent can't engage in *negative* contributions – assault, theft, or other rights violations, say. Taking absence as the baseline condition amounts to assuming a precontractual state of nature in which no violations of (negative) rights occur.[207] That is, it amounts to a Lockean baseline. Absence is precisely the counterfactual Gauthier uses.[208] As I have already argued, and will do so in greater detail later, a Lockean baseline is not available to a contractarian theory.[209] Contractarian theories must instead be committed to a Hobbesian baseline, at which agents maximally employ their capacities for force and fraud to secure as advantageous a non-agreement point for

[205] 2016: §5.3.2, emphasis added.
[206] *Ibid*.: §5.3.2 n17.
[207] Muldoon in places seems to be assuming that agents would have a robust capacity to exit societies which do not satisfy his proportional gain principle. I see no reason to assume this.
[208] 1986: 200-205. See §7.5, *infra*.
[209] §1.6.3, §7.5.

themselves as possible.

I will ultimately argue, with Muldoon, that an analysis of coalition dynamics makes a crucial contribution to a successful contractarian theory. I will aim to repair what strike me as the infirmities in his development of this idea, and hence to put Muldoon's insight on a firmer analytical foundation.

§1.7 A contractarian conception of justice

In political social contract theories, the object of agreement is a set of normative principles to regulate the political interaction of the people to whom the theory is applied. Agents representing idealised versions of these people are placed in a hypothetical choice situation and must agree on the principles selected. Famously, in Rawls's theory, we can say that whatever principles would be chosen in the hypothetical situation are the true and correct principles of justice.[210]

Rawls is interested in defending a particular conception of justice. He begins by distinguishing between a conception of justice and the *concept* of justice – what a conception of justice is a conception *of*.[211] By way of analogy, we can conceive of different kinds of chairs, but there must be some kind of link among the various examples in order for us to say that any particular one is a conception of a chair, rather than a conception of an elephant. The concept of a chair links the various conceptions of chairs, and will in part be defined functionally – chairs are objects we use for sitting.

According to Rawls, the concept of justice is also defined functionally. Justice plays a particular role in social organisation. Rawls supposes that we find ourselves in circumstances where we can improve our lot if we coordinate our actions with one another. The benefits of coordination, and the burdens of producing them, can be

[210] 1999a: 3-15; but compare Rawls, 2005: 94, 116, 126-129.
[211] Compare Vanderschraaf, 2019: 272.

distributed in different ways. Principles of justice are principles designed for choosing among the various possible social arrangements which produce these different distributions of burdens and benefits.[212] Thus, a principle which said that the benefits of coordination (morally-)ought to be distributed in proportion to how tall the participants are would be a principle of justice. It would, of course, be an implausible principle of justice, but we can distinguish between the form of a principle and its content. The height principle has the right form – it is a principle which selects a particular distribution of burdens and benefits of coordination – and so it is a conception of the concept of justice. It has the wrong content, however, because it is not *true* that the burdens and benefits of coordination (morally-)ought to be distributed according to height. Thus, while the height principle is a conception of justice, it is not the correct conception of justice.[213]

Principles of justice are clearly normative on Rawls's account – they tell us how something (in-some-sense-)ought to be distributed. Although he is more explicit in his later than his earlier work, the normativity Rawls has in mind is moral.[214] Thus, for Rawls, a conception of justice tells us how the burdens and benefits of coordination (morally-)ought to be distributed. A political contractarian account has the same structure. The object of agreement is a set of normative principles to regulate political interaction, and so to distribute the burdens and benefits of coordination. However, the normativity in question is prudential rather than moral. In my account, the principles selected in the hypothetical choice situation will tell us how the burdens and benefits of coordination (prudentially-)ought to be distributed.

Insofar as we think that justice is essentially a moral concept, there cannot be a political contractarian conception of justice.[215] Nonetheless, I am aiming at a prudential analogue for justice. The only difference between the analogue and the

[212] Rawls, 1999a: 4-6; 109. Compare Barry, 1995: 72; Vanderschraaf, 2019: xii, 67, 89.
[213] Rawls, 1999a: 9. Compare Korsgaard, 1996: 113-114.
[214] 2005: 484 n91; 2001: 6-7. See also the discussion of 'to each according to his threat advantage' in §3.2.2, *infra*.
[215] However, see the discussion of prudential and moral reasons in §2.4. Compare Barry, 1995: 39-45.

Rawlsian concept will be the kind of normativity underlying each.

When justice is discussed from the point of view of moral contractarianism, the term used to describe its conception of justice is 'justice as mutual advantage'.[216] Because moral contractarians understand moral normativity to be derived from prudential normativity, there is a sense in which their conception can be moral while still ultimately being grounded in prudential normativity, and so can still genuinely be a conception of justice.[217] That sense is not available to the political contractarian, but I propose to borrow the term 'justice as mutual advantage' to refer to the contractarian analogue anyway. Despite the fact that the description is ironic from the point of view of political contractarianism, I lack a better label. With this caveat in mind, what follows is an account of justice as mutual advantage.

[216] Barry, 1989: 5-8; Barry, 1995: 32; Vanderschraaf, 2019: 273-274.
[217] Gauthier, 1986: 4-5, 150-156, 233-267; Vandershraaf, 2019: xi-xii. Compare n14, *supra*.

Chapter 2

Political contractarianism claims that our prudential reasons should guide us when we are thinking about how to organise our political lives. This is not because the theory denies the force of our moral reasons, or is sceptical about their existence. On the contrary, it allows that one of our most important goals could be the pursuit of what we take to be justice. Instead, reliance on prudential reasons is a response to the problem of moral disagreement.

In this chapter, I argue for political contractarianism. In §2.1, I defend a conception of politics. I argue that, among other things, politics involves giving one another reasons of various kinds. In §2.2, I distinguish two different levels at which reasons operate in my argument. At the first level, there are the kinds of reasons we use to justify political actions to one another. At the second level, there are the reasons we have which justify us using particular kinds of reasons at the first level. My account tells us to use prudential reasons at the first level, and gives each of us reasons at the second level to justify us using prudential reasons at the first level. To mark the distinction, I will call reasons at the first level '*u*-type' reasons, and reasons at the second level '*i*-type' reasons.

In §2.3, I argue that, given my conception of politics, we have an *i*-type prudential reason to justify political actions using *u*-type prudential reasons. That is to say, I argue that we (prudentially-)ought to be political contractarians. This still leaves open the question of whether we (all-things-considered-)ought to be political contractarians. In §2.4, I argue for the all-things-considered claim.

The discussion of reasons in this chapter is extensive, so it will be helpful to keep the ultimate payoff in view. Political contractarianism is a social contract theory. In such a theory, the *basis of agreement* is whatever *u*-type reasons the theory

endorses.[1] *i*-type reasons justify using a particular set of *u*-type reasons. In Part II, I construct a contractarian model to justify liberal political institutions. The upshot of *this* chapter is a justification for using a model of *that* kind – one which relies solely on prudential reasons – to select political institutions, rather than a model of some other kind.

§2.1 The scope of politics

I define a political action as any action that is intended by an agent to have an effect (direct or indirect) on the intended actions of some other agent.[2] That is, I take a political action to be any action that is *strategic*. The political domain is constituted by the set of possible strategic actions.[3]

This is a broad definition of politics. It will be far too broad for some, and I do not have the space to defend it fully. Instead, I will argue that political actions have at least three necessary features: they are necessarily *social*, *purposive*, and *deliberative*. I think these three features are also jointly sufficient to define politics, but my argument will not rely on the sufficiency claim, and so for the purpose of this argument I can remain neutral with respect to any further conditions which might be needed to characterise politics. In the remainder of the chapter, I will argue that any action meeting these three necessary conditions is properly guided by prudential reasons. So long as all political actions meet the conditions, my argument will go through, even if the three conditions are not sufficient to characterise politics.

To begin, politics is clearly *social*. Politics is not something I can do by myself, and it is not something which can occur between me and a rock. It is something that

[1] See the introductory remarks in Chapter 1.
[2] Clark, Golder, and Golder, 2013: 50, 52. Compare the discussion in Rehfeld (2010: 472-473), citing Dahl (2004: 377) and Freeden (2005: 116). Rehfeld suggests that politics necessarily involves the exercise of power by one agent over another; Dahl refers to power or influence. Rawls's understanding of the political domain is narrower (2005: 135-137).
[3] Compare Vanderschraaf, 2019: xiii, 43.

occurs when agents interact in some way.[4] However, politics doesn't include all possible interactions between agents. A conversation about the weather, or about conceptions of justice, is not politics. A conversation about monetary policy or abortion laws might be a conversation *about* politics, but it does not itself constitute politics.[5]

Second, I take politics to be *purposive*. That is, I take it to be an intentional, goal-directed activity.[6] Human beings are entities with goals. There are states of affairs we seek to bring about, and other states of affairs we seek to prevent from coming about. Some of our goals we can accomplish by ourselves, just by rearranging bits of the material world (including the bits of the material world which constitute our own bodies). Other goals can be accomplished only if the people around us act (or refrain from acting) in particular ways. Being social, politics is concerned with goals of the latter kind. It is the activity that individuals undertake with the intention of shaping the behaviour of other people in a manner conducive to the realisation of those goals. Thus, in classical social contract theories, individuals choose to enter political society with one another in order to escape the perils of the state of nature. In contemporary democratic politics, we vote in an attempt to control the goals the state will promote, and we campaign in order to convince others to lend their support to our efforts. In each case, our interaction with others is directed at a purpose.

Third, politics is *deliberative*. By this I mean it is concerned with influencing how people *choose* to act.[7] The relevant contrast is with directly controlling people via brute force. For example, suppose I push you off a bridge. This action is social, as it involves an interaction between two agents. It is also purposive – we can suppose

[4] *Ibid*.: 3, 43.
[5] McDermott, 2008: 25; Rehfeld, 2010: 476. McDermott and Rehfeld are concerned to define political philosophy and political science respectively, but the way in which each distinguishes the *study* of political phenomena from the *practice* of politics itself helps to illuminate what politics is.
[6] Compare Vanderschraaf, 2019: 55.
[7] Compare Leys, 1965: 274. Strictly, any action that is deliberative is also social, and so there is no need to specify these as independent conditions. I do so for ease of exposition – I take it the deliberative condition is more controversial, and it is easier to begin with the more obvious claim that politics is social. I thank Robert Jubb for pressing this point.

that I intend to bring it about that you fall in order to advance one of my goals. Nonetheless, our interaction is not political. In pushing you, I am not attempting to bring it about that you decide to act one way rather than another, even though I am intending that your body moves in a certain way. My use of force is an *alternative* to politics. In the bridge case, I can accomplish my goal by directly rearranging certain parts of the world (your body and its position relative to the bridge), and so I have no need for politics. More generally, I am not engaging in politics when I interact with others in a manner that aims to bypass their deliberative faculties entirely.[8]

This is not to say that a political action can never involve the use of force or violence. Suppose my purpose in pushing you from the bridge is *pour encourager les autres* – my aim is to communicate a message to some audience, and I am using your body as a means of sending the message. This would qualify as political on my account. It would be a social interaction between me and my audience, and it would be purposive. It would be an attempt to influence how my audience decides to act, and hence it would also be deliberative. You and I would not be interacting politically – from my point of view, you might as well be a rock – but my action would be political.[9]

What does it mean to engage or bypass a person's deliberative faculties? Consider an inflammatory public speech, designed to whip a crowd of people into an emotional frenzy, delivered with the intention that they will act on the basis of those feelings toward some goal at which the speaker aims. The act of making the speech is certainly social and purposive. It also seems quite natural to see it as a political action. Has the speaker engaged the deliberative faculties of the audience? If not, and if we would still consider making the speech a political action, should we reject the claim that politics is necessarily deliberative?

I take it that a person's deliberative faculties are engaged when, and only when, she

[8] For an alternative perspective on this question, including a discussion of deception, see Horton (2010: 438-440, 442-443).
[9] I thank Philip Smith for helpful discussion of this point.

decides to act on the basis of her reasons. There are circumstances in which people do not decide to act on the basis of their reasons – when they react without thinking, or somehow lose control of their behaviour. If I aim to bring it about that a person does not decide to act on the basis of her reasons, I am aiming to bypass her deliberative faculties. By contrast, aiming to engage a person's deliberative faculties means providing her with reasons upon which she can decide to act.[10]

In the case of the inflammatory speech, if it really is true that the people in the crowd are overcome by emotion, and react in a non-deliberative fashion, then we should think of this case as akin to a case of brute force. That is to say, there is no relevant difference between me pushing you off a bridge, and you leaping from the bridge in uncontrollable panic because I have terrified you. In both cases, by hypothesis, you lack control of your movement. As such, the inflammatory speech would not be a political action. Lest this seem counter-intuitive, I would add that it strikes me as unlikely that a speech could have such an effect. This is not to say that such things never happen (much less that they are never intended), just that most of the time a speech of this kind will not *compel* us to act, in a way that completely bypasses our deliberative faculties. And, insofar as the speech simply attempts to change the way in which we order our reasons, or to bring to our apprehension reasons which we had not previously considered, it counts as deliberative. Thus, attempts to control people through inducing rage or terror, or through something more exotic like brainwashing or hypnosis, fall outside the political domain in the same way that the use of brute force does.

The case of manipulation or deception is harder to judge.[11] In moral and legal discourse, force and fraud are close cousins. Both are thought to subvert the autonomy of their target. Force is excluded from my conception of politics because it bypasses a person's deliberative faculties. There is a sense in which fraud also

[10] For further discussion of what it means to act on the basis of a reason, see §2.3.1 (compare also the discussion in Enoch, 2011: 221-233). It should be clear that I have a broad understanding of acting on the basis of a reason. In particular, there is no necessary inconsistency between being motivated by emotion and acting on the basis of a reason.
[11] I distinguish, here, between manipulation and persuasion (Dahl and Stinebrickner, 2003: 40-41).

bypasses our deliberative faculties. It seeks to deny us access to the genuine reasons we have for action. However, there is sense in which it does not bypass our deliberative faculties entirely, since it generally seeks to substitute false reasons upon which those faculties can operate.

It seems to me that too much of what we ordinarily think of as politics would be excluded from the concept if manipulation and deception were to count as bypassing a person's deliberative faculties. As such, we should say that I am acting politically whenever I attempt to influence your behaviour by giving you something I intend you to take as a reason to decide to act in a particular way. I will describe this latter sense of a reason as a *subjective* reason.[12] The possibility of both sincere and manipulative political actions is thus left open.

Because I take politics to be deliberative, I take it necessarily to be concerned with giving people reasons. There are several ways in which to understand what it means for me to 'give' you a reason. Most straightforwardly, it could mean that I engage you in conversation and perform speech acts of a particular kind. So, saying "If you mow my lawn, I'll pay you £10" is an example of giving you a reason to act in a certain way (in a way consistent with mowing my lawn, assuming you value money, that my offer is credible, and that we are not engaged in something like a play or performance). However, we can also understand reason giving in a wider sense. When I cut my prices, I intend that customers will choose to buy my goods rather than spend their money elsewhere; no explicit conversation has taken place, but my intention is to give prospective customers a reason to shop at my store which they did not have before, and so to influence how they choose to act. I will take it that politics can involve giving people reasons either explicitly or implicitly.

Nonetheless, not all cases of reason giving will count as political actions on my account. If I mention to you that the local garden centre is having a sale on lawn mowers, there is a sense in which I have given you a reason to purchase a mower.

[12] See §2.3.1.

However, unless I did so purposively – with the intention of getting you to act in a particular way – I will not have acted politically. Similarly, there is a sense in which I give you a reason to mug me if I walk down the street, whistling and tossing my cash-stuffed wallet in the air. But, again, this wouldn't count as a political action on my part, because it is not my intention to influence your decisions. All political actions will involve reason giving, but not all reason giving is political.

§2.2 Two levels of political reasoning

Politics is an activity in which we aim to influence how other people decide to act by giving them reasons. Political contractarianism is the proposition that, when engaging in politics, we ought to give others prudential reasons rather than moral reasons. My argument is that we ought to be political contractarians.[13] That is, I argue that we have a reason to give other people reasons of a certain kind – prudential reasons – when engaging in politics.

What we are talking about here is a reason to give reasons. To avoid confusion, it will be important to keep the following distinction clear. It is one thing to say that, when acting in the political domain, one has a reason of some type, *t*, to give other agents prudential (or, alternatively, moral) reasons. It is another thing to say that, when acting in the political domain, one has a prudential (or, alternatively, moral) reason to give other agents reasons of type *t*. Reasons enter the discussion at these two different levels.

Let me introduce some terminology to mark this distinction. Suppose that *φ*-ing consists of giving you a reason. I want to distinguish between (i) the reason I have to *φ* (call this an '*i*-type' reason, because it is the reason by which *I* am justified in *φ*-ing) and (ii) the content of what I give you when *φ*-ing (call this a '*u*-type' reason,

[13] For the time being, I leave these 'oughts' unqualified, because the purpose of this section is to describe a system for qualifying them.

because it is the reason which is supposed to justify *you* acting in some way).[14] My argument is then that we have an *i*-type reason to employ prudential *u*-type reasons when justifying political actions to one another.

The basic thought underlying the *i*/*u* distinction should be quite familiar. What I have described is a general structure for a theory of 'public reason', made up of three components:[15]

> First, '*φ*-ing' consists of reason giving (which is to say, action) in a political domain.[16]
>
> Second, '*u*-type reasons' make up the content of political intercourse. These are the reasons which pass the bar of 'public reason', however that is defined.[17]
>
> Third, '*i*-type reasons' define the bar of public reason. These are the reasons which justify the inclusion or exclusion of particular kinds of *u*-type reasons in or from political intercourse.

Depending on how we fill in these components, we can generate a range of conceptions of public reason.[18]

The point upon which I want to focus here is this: *i*-type reasons are not necessarily moral reasons. They might also be prudential reasons.[19] I could have either prudential reasons or moral reasons to offer you reasons of a particular kind when acting in the political domain. This being the case, either moral or prudential reasons

[14] Rossi (2010: 29, particularly n22) also articulates something like this distinction.
[15] For discussions of the concept of public reason see, among many others, Rawls (1997; 2005: 212-254), Gauthier (1995), Larmore (2003), Chambers (2010), Gaus (2011: 24-36), Quong (2011: 256-289), Thrasher (2019: 430, 432).
[16] §2.1. Compare Rawls, 1997: 767-768; Rawls, 2005: 136, 137.
[17] Chambers, 2010: 894. Compare Rawls, 2005: 220 n7; Quong, 2011: 173.
[18] Compare Rawls, 1997: 795. On the possibility of a range of conceptions of public reason, see *ibid*.: 772.
[19] The same, of course, holds for *u*-type reasons, but my focus in this section is on *i*-type reasons. For simplicity, I have restricted my discussion to prudential and moral reasons. See §1.3 n12.

can define a bar of public reason. Rawls's theory of public reason – the theory with which many of us are most familiar – consists of a moral *i*-type reason (the liberal principle of legitimacy, and the associated duty of civility) which admits into political intercourse a variety of moral and prudential *u*-type reasons.[20] But moral reasons are not the only way in which to specify a theory of public reason.

By way of illustration, I could have an *i*-type moral reason to give other agents *u*-type moral reasons. I might have a moral duty (an *i*-type reason), when employing the coercive power of the state, to give those over whom I am exercising the power a moral justification (a *u*-type reason) for complying with my instructions. This duty (the *i*-type reason) might ultimately be grounded in respect for my fellow citizens as moral agents.[21]

But I might also have an *i*-type moral reason to give another agent a *u*-type prudential reason for taking a particular action. If I am an armed police officer, and I come across one agent violently assaulting another, I have a moral duty (an *i*-type reason) to draw my weapon and threaten to shoot the attacker unless she ceases her attack and surrenders. Threats are a way of giving other agents prudential reasons to decide to act in compliance with the threat. A police officer's threat is a *u*-type prudential reason, the giving of which is justified by an *i*-type moral reason.[22]

Similarly, I can have an *i*-type prudential reason to give other agents *u*-type prudential reasons. If threatening to shoot you will result in you deciding to give me your money, then I have an *i*-type prudential reason to threaten to shoot you (assuming I want your money). Threatening you is a way of giving you a *u*-type prudential reason to comply with my demand. And so, in this case, I have an *i*-type prudential reason (because issuing the threat is an effective means to my end) to give

[20] Rawls, 2005: 137, 217; Rawls, 1997: 768-769, 770-771; Quong, 2011: 256-257.
[21] Rawls, 1999a: 156, 297; Larmore, 2003: 373.
[22] The extent to which liberal political philosophers are committed to using *u*-type prudential reasons is sometimes obscured by their assumption of strict compliance. Focusing on ideal theory typically obviates the need to consider what (morally-)ought to be done in the face of non-compliance with (what one believes to be) justice. See Rawls, 1999a: 8, 125-126, 215-217, 308-309; Farrelly, 2007; Stemplowska, 2008.

you a *u*-type prudential reason (the threat).

Less obviously, I can have an *i*-type prudential reason to give other agents *u*-type moral reasons. If I am starving, and you have food, I could inform you that you have a moral duty to share the food with me. Insofar as the moral duty would move you to act if I made you aware of it (that is, insofar as the moral reason I could give you would be *effective*), I have an *i*-type prudential reason to give you a *u*-type moral reason, in order to accomplish one of my goals (avoiding starvation).[23]

The qualification "insofar as the moral reason I could give you would be effective" exposes the deeper structure of the argument I am making. Whenever I give you a reason for the purpose of accomplishing one of my goals (and, by definition, this includes all political cases), I have an *i*-type prudential reason to give you a reason that is effective in bringing it about that you decide to act in the way I suggest. This follows from the nature of prudential reason.[24] Conceiving of politics as a purposive, goal-directed domain necessarily brings it into this relationship with prudential reasons and effectiveness.

§2.3 Reasons and motivation

With a conception of politics and the *i*/*u* distinction on hand, I can now present my argument for political contractarianism. For the sake of clarity, I will first present a simpler, overstated version of the argument; I will then add some necessary caveats. To begin:

> Politics is a purposive activity. When one is engaged in a purposive activity, one has an *i*-type prudential reason to select effective means to one's ends. Politics is also a deliberative activity. When one is engaged in politics, what

[23] I discuss effective reasons in more detail in §2.3. It should be clear from this last example that saying I have a prudential reason to φ is consistent with saying that I have a moral reason to φ.
[24] Compare Rawls, 2005: 50-51, 83, 165.

one is doing is giving other people *u*-type reasons to decide to act one way rather than another. Therefore, when one is engaged in politics, one has an *i*-type prudential reason to give others effective *u*-type reasons.

Because of the problem of moral disagreement, moral reasons are not effective in politics. Conversely, prudential reasons are effective. Therefore, when one is engaged in politics, one has an *i*-type prudential reason to give others *u*-type prudential reasons.

Political contractarianism is the proposition that, when one is engaged in politics, one (in-some-sense-)ought to give others (*u*-type) prudential reasons. Therefore, one (prudentially-)ought to be a political contractarian.

I will call this the *argument from effectiveness*. As I noted, this version of it is overstated. It is too much to say that moral reasons can never be effective, or that prudential reasons are always effective. Nonetheless, in outline, this is my argument for political contractarianism. In §2.3.1, I discuss effective reasons. In §2.3.2, I defend the claim that moral reasons (typically) won't be effective in a political context, but that prudential reasons (typically) can be. In §2.3.3, I consider cases in which moral reasons are effective, and cases in which prudential reasons cannot be.

§2.3.1 Effective reasons

Given that politics is purposive and deliberative, I take it to be uncontroversial that when one is engaged in politics, one has an *i*-type prudential reason to give others effective *u*-type reasons. What, then, is an effective reason?

Consider an example of reason giving: Suppose Willow tells Xander that he (morally-)ought not to steal. When Xander asks why, Willow tells him that stealing is morally wrong. Willow will have given Xander an effective reason if the consideration she advances ("stealing is morally wrong") is, all else equal, sufficient to bring it about

that Xander decides to do as Willow says (refrain from stealing).

Willow can fail to give Xander an effective reason in a number of ways. First, Willow's utterance could fail to be a reason. Suppose that, in place of saying that stealing is morally wrong, Willow were to say, "Abracadabra!" Suppose further that Xander has been hypnotised, such that he refrains from stealing whenever he hears someone say, "Abracadabra!" In this event, while Willow's utterance is effective in preventing Xander from stealing, it is not an example of her giving Xander a reason to refrain from stealing (and so it is not an example of her giving Xander an effective reason).[25] This is true even if the hypnotically implanted phrase is "stealing is morally wrong". In order for Willow to give Xander a reason, she must do more than perform a speech act which has the consequence of influencing his behaviour – she must engage Xander's deliberative faculties.[26]

Engaging Xander's deliberative faculties means presenting him, explicitly or implicitly, with an argument. Thus, Willow's utterance must amount to something like the following:

P1: Xander (morally-)ought not to do things which are morally wrong.

P2: Stealing is morally wrong.

Therefore,

[25] Why isn't "Abracadabra!" a reason to refrain from stealing? It is beyond my scope to provide a full account of what a reason is, but, very roughly, I take a reason to be a consideration in favour of one action over some other(s) (Vallier, 2016: 599). At the least, then, a reason must provide us with some way of distinguishing between actions the reason prefers and disprefers. "Abracadabra!" fails on this criterion. However, suppose Willow were to say, "Because stealing is not cool." This utterance does distinguish the two possible actions. It does so relative to a (putatively) normative standard of cool actions. But Xander will only understand the utterance as a reason if he comes from a certain cultural background. What can be recognised as even a putative reason will be to some extent culturally specific – Kant probably wouldn't understand a Californian appeal to cool; Xander wouldn't understand a reason offered in German. For further discussion of some related complexities, see Enoch (2011: 221-233) and Vallier (2016).
[26] §2.1.

C: Xander (morally-)ought not to steal.

If Xander believes that he (morally-)ought not to do things which are morally wrong; and if Willow's utterance leads him to believe that stealing is morally wrong; and if Xander can reason sufficiently well; then Xander will necessarily believe that he (morally-)ought not to steal.[27] In this event, Willow will have given Xander an effective reason. She will have given him a reason, because she will have engaged his deliberative faculties in the right kind of way – the belief that he (morally-)ought not to steal will be a conclusion Xander reaches via his own reasoning. And the reason will be effective, in that it will (all else equal) result in Xander deciding not to steal.

Taking this model of deliberation, we can identify two further ways in which Willow can fail to give Xander an effective reason. First, Xander might not believe one of Willow's premises. In this case, the reason Willow gives Xander is ineffective because it implicitly relies on the truth of a proposition he believes to be false. Second, Xander's reasoning may fail, such that he does not validly conclude that he (morally-)ought not to steal, despite accepting Willow's premises. If this looks odd in so simple an example, consider the complexity of the reasoning which could be involved in political arguments for tax policy or criminal justice regimes.

This account of effective reasons maintains that whether or not a reason is effective depends on the beliefs of the agent to whom the reason is given. Note that this does not amount to saying that Willow has not given Xander a reason to refrain from stealing if he does not believe her to have given him a reason. *P2* can be true – and so Xander really does have a reason to refrain from stealing, whatever his beliefs – without it being an effective reason (when given to Xander). Conversely, if *P2* were false, but Xander believed to it be true, it would be an effective reason (when given to Xander) despite not being true. Similarly, what is important in determining the

[27] See Broome's discussion of normative requirement (1999: 408-409; but see also §5.1 n8, *infra*). If Xander is rational, as I will understand that term, it follows from this belief that (all else equal) he will decide to act on it (§5.1, *infra*). For a discussion of the assumption that the people with whom we interact politically are (more or less) rational, see §7.3, *infra*.

effectiveness of the reason is that Xander believes the implicit argument to be valid, not that it actually is valid.[28]

The upshot is that, in order to give someone an effective reason, we must appeal to beliefs she already holds – we must reason to a normative conclusion using only premises which our target believes to be true, and by methods she can understand. To put this another way, to be effective a reason must be a *subjective* reason; it is neither necessary nor sufficient that the reason is an objective reason.[29]

None of this is to say that our objective reasons are unimportant. Ideally, we would want our subjective reasons and our objective reasons to overlap perfectly – we would want to have perfectly accurate beliefs about our objective reasons. We often deliberate carefully in order to achieve this. However, while we can act *in accordance with* our objective reasons, we can only ever act *on the basis of* our subjective reasons, in the hope that our subjective reasons track our objective reasons as closely as possible. Since politics is concerned with giving others (*u*-type) reasons upon which they can decide to act, it must be concerned with giving people subjective reasons. Only subjective reasons can be effective reasons, because only subjective reasons are connected in the right way to the deliberative process by which an agent decides to act.

§2.3.2 Pervasive and persistent disagreement

In circumstances where there is moral disagreement – in the circumstances in which

[28] Again, see §2.1, particularly the discussion of fraud and manipulation. I thank Robert Jubb for discussion of these points.

[29] See the discussion of idealisation in the Interface. On the distinction between subjective and objective reasons, see Lenman (2009: §1; compare Korsgaard, 1996: 134-135, 149-150; Schroeder, 2018: 292-295). Subjective reasons are not necessarily internal reasons. In order to have an internal reason, there must be a sound deliberative route from an agent's existing commitments to the reason, but it is not necessary that the route be subjectively available to the agent (Williams, 1981: 101-113; Chappell, 2013: §4; compare Korsgaard, 1996: 82 n63). However, this distinction is less important in light of the idealising assumptions about reasoning and information that I later make in developing the contractarian model (Part II, *infra*). Note also that a subjective reason is not simply an explanatory reason; it is (also) a justifying reason (Lenman, *op. cit.*; and §2.4.1, *infra*).

citizens of contemporary societies find themselves – appeals to moral reasons will often fail to be effective. They will often rest, in the end, on an appeal to a fundamental moral principle which is not shared by the fellow citizens to whom one is addressing the appeal. By contrast, prudential reasons – reasons which justify a prescription on the basis of advancing the goals of the agents to whom they are addressed – will always be subjective, and hence, in principle, can always be effective.[30] That is why prudential reasons are the appropriate constituent elements of politics, and why moral reasons are often not.

This is the argument from effectiveness. It relies on a claim about the extent of disagreement on moral principles. But, of course, moral disagreement is not the only kind of disagreement there is.[31] On the account I have given, any kind of disagreement can render a reason ineffective, since any disagreement about the truth of a premise upon which a prescription implicitly relies will eliminate the force of the prescription from the point of view of the dissenting agent. So, why focus on *moral* disagreement? For two reasons: moral disagreement is often (i) pervasive and (ii) persistent.

First, the problem of *pervasiveness*. There is widespread disagreement about which flavour of ice cream is most delicious. Any reason given in the course of a political interaction which relied on the premise that vanilla is the most delicious flavour would be ineffective for a significant number of people. Nevertheless, it is difficult to see how very many such reasons could plausibly rely on the vanilla premise. Similarly, there is significant disagreement among physicists about the correct interpretation of quantum mechanics, but not many people are physicists. Very few people make any decisions which turn on adopting a particular interpretation of quantum mechanics.[32]

---

[30] Whether or not they are effective in any particular case will depend on the precise structure of the preferences of the agents concerned. See §2.3.3 and §5.5.2.

[31] Rawls, for example, recognises that insofar as there is disagreement (presumably reasonable, though he doesn't quite say so) about non-moral propositions, public reason also excludes claims which rely on those controversial propositions (2005: 224). Compare Abizadeh, 2011: 309.

[32] My claims, here, refer to the world in which we actually live. We can imagine possible worlds in which disagreement about ice cream or quantum mechanics is of much greater consequence.

By contrast, moral propositions can be implicated in a vast number of the decisions we make, and (at least if common moral beliefs are true) are relevant to decisions made by just about any person. Take, for example, the proposition that each person is of equal moral worth. This proposition might plausibly appear in (ultimate) justifications for any number of actions I could take in the course of a day. If one is inclined to doubt this, consider that many of our decisions are decisions not to act (for example, in ways that violate the rights of others) or are decisions to act which carry opportunity costs (spending money on a cinema ticket, rather than donating to a charity).

Disagreement about ice cream and quantum mechanics is not pervasive, because its scope for rendering reasons ineffective is very limited. Propositions about ice cream and quantum mechanics are highly unlikely to appear among the premises of the implicit argument backing a reason. Moral disagreement, on the other hand, is pervasive both in the sense that it potentially touches a wide range of decisions and in the sense that it touches decisions made by most people.

Now, one might worry that the contrast I have drawn has less to do with the difference between moral disagreement and other kinds of disagreement, and more to do with the level of generality at which the examples have been specified. The moral equality of persons is a moral proposition expressed at a high level of generality, whereas the ice cream and quantum mechanics cases are more specific. It is hardly surprising that very general propositions are pervasive. If, instead, we take a more specific moral proposition, we will find it is no more pervasive than the non-moral propositions I have mentioned. Consider, for instance, the claim that I (morally-)ought to leave a tip in cash at a restaurant if I am not sure that tips left electronically will be distributed appropriately to the staff. This is a moral proposition, but it is hard to see how it could be relevant to very many of my actions.

This point is well taken, but it does not undermine the claim that moral disagreement is particularly likely to be pervasive. If the tips premise is appearing in an implicit

argument, it will in turn almost certainly rely implicitly on a more fundamental moral principle (a duty to treat people fairly, say). To the extent that any moral proposition which is not pervasive implicitly relies on a more fundamental moral proposition which is pervasive, any ineffectiveness in the latter will infect the former. Thus, moving to more specific moral propositions does not make those propositions any less likely to be ineffective.

One could resist this conclusion in two ways. One might endorse moral particularism, and thus deny that non-pervasive moral propositions will implicitly rely on pervasive moral propositions. But this would be self-defeating. The truth of the non-pervasive propositions would then rely on the truth of moral particularism, and the latter is a pervasive proposition about which there is extensive disagreement.

Alternatively, then, one might say that non-moral propositions are in the same boat as moral propositions – the specific proposition that vanilla is the most delicious flavour of ice cream implicitly relies on something much more general, like the claim that there exists an external world which contains different flavours of ice cream. This is true, but in order for it to justify us treating moral disagreement as no different from other kinds of disagreement we would have to think that disagreement about very general non-moral propositions is about as likely as disagreement about very general moral propositions to result in political prescriptions being ineffective. And that seems radically implausible. There are people who deny things like the existence of mind-independent reality, or inductive knowledge, but the vast majority of our fellow citizens will not dispute these general premises, while many of them will dispute general moral premises.

Thus, there really is a difference between the pervasiveness of moral disagreement and the pervasiveness of other kinds of disagreement. The difference isn't simply an artefact of cherry picking specific rather than general examples. This is not to say that other kinds of disagreement can never lead to a reason being ineffective, simply that focusing on moral disagreement in particular is warranted.

The second reason to focus specifically on moral disagreement is the problem of *persistence*. It is a commonplace in discussion of moral disagreement to note that it cannot always be resolved through rational argument or debate. Attempts to resolve the disagreement often seem to run into a regress problem. If I assert some moral proposition which you do not already endorse, you might reasonably ask me to justify the proposition. Typically, I will at that point appeal to a more general moral proposition in support of my initial claim. But, if there is also disagreement about that more general proposition, I will not have given you a subjective reason to accept my initial claim. I can attempt to retreat further up the chain of justification, but sooner or later I will reach the end of my chain of reasons – some proposition I take to be true, but for which I cannot offer you any justification. If the disagreement between us extends to this fundamental level, I cannot give you a (subjective) reason to accept my initial claim.[33] Call this *fundamental* disagreement.

Cases of fundamental disagreement can arise with respect to any kind of disagreement, but they seem likely to arise with particular frequency in moral disputes. People endorse a wide variety of fundamental moral principles, for which the only support that can be given is a blunt appeal to intuition. By contrast, disagreements about things like empirical facts or scientific theories, while certainly real, do not generally display this kind of intractability.[34] We may not be able to agree on the facts themselves, but we can often agree on shared standards for establishing the truth of the matter – we can gather evidence, and we can test theories, following a shared methodology. In constructing a social contract theory, we have the luxury of assuming that the participating agents have taken the time to work through the methods of the best available science, for example, in a way that real political agents often will not.[35] We seem to lack equivalent methodological tools in the moral realm; or, at least, the methods we have for settling moral

[33] Compare Alston, 1976: 178; Barry, 1995: 28-31; Abizadeh, 2011: 309. For ease of exposition, I have assumed a foundationalist structure for moral beliefs. But the situation is not improved by adopting a coherentist approach, say.
[34] See Rawls, 2005: 55. Aesthetic disagreements, or simple disagreements of taste (like the ice cream case), are likely to be persistent in the same way that moral disagreements are, but my sense is that they are much less likely to be pervasive.
[35] However, see the discussion of idealisation in the Interface.

disagreements are themselves at least as controversial as the substantive moral propositions they purport to test.[36] And so, we should not suppose that simply allowing the agents more time, say, to study the problem would necessarily resolve disagreement.

Given the nature of moral disagreement in the world in which we find ourselves, appeals to moral reasons will often result in a clash of fundamental moral commitments that is not resolvable through rational argument – disagreement is persistent, even in the face of good faith efforts to resolve it. In such cases, the moral reasons given to a dissenting agent cannot be subjective reasons (though they may well be genuine objective reasons, in the event that the dissenting agent is mistaken about the moral facts of the matter). Since they cannot be subjective reasons, they cannot be effective reasons, and so they cannot be the kinds of reasons which are appropriately given in politics.

All of this would be beside the point if prudential reasons were equally likely to be ineffective. However, they do not generally give rise to the same problems. Consider a prudential version of the example presented earlier:

P1: Xander (prudentially-)ought not to do things which result in him being shot.

P2: Willow will shoot Xander if he steals.

Therefore,

C: Xander (prudentially-)ought not to steal.

[36] Moreover, they are so controversial for exactly the reason to which Dworkin appeals in rejecting moral scepticism (1996: 118; see §1.5, *supra*). Different methodologies give rise to different moral prescriptions, but it is the moral prescriptions themselves to which we are often committed in the end. Compare Rawls: "Political philosophy cannot coerce our considered convictions any more than principles of logic can... It is a mistake to think of abstract conceptions and general principles as always overriding our more particular judgements." (2005: 45; see also, *ibid*.: 242 n31.) It is hard to imagine any sensible person making a similar claim about, for example, the scientific method.

It should be clear that, insofar as Willow's threat is credible, and Xander prefers not being shot, the threat gives Xander a subjective, effective reason to refrain from stealing. Disagreement is still possible here, of course. Xander may believe *P2* to be false. Such disagreement is unlikely to be persistent, however. Willow and Xander can test the truth of *P2* quite straightforwardly.

Nonetheless, not all cases involving prudential reasons will be so simple. Suppose Willow were to try to convince Xander to quit smoking by telling him that, unless he does so, he will be much more likely to suffer from cancer in the future. It is not too hard to imagine a person for whom this reason would not be effective.[37] Quite apart from the fact that there are people who reject the scientific consensus linking smoking and cancer, and even reject all reasonable attempts to demonstrate the link by way of standard methodologies, there is still the uncertainty as to whether or not any particular agent will suffer as a result of smoking. Such uncertainty will inevitably be a feature of arguments grounded in consequences which occur in the relatively distant future. Even though Willow has offered Xander a prudential reason to quit smoking, the reason might not be effective.

What the smoking case establishes is that not all prudential reasons are effective. However, political contractarianism does not rely on the claim that *any* prudential reason will be effective. My argument is that, in the context of contemporary societies, moral reasons will typically be ineffective for a significant number of the fellow citizens with whom one is interacting politically, but that there will often exist some prudential reasons which can be effective. The existence of prudential reasons which are ineffective does not undermine that claim.

[37] It is harder to imagine this obtaining in the more idealised circumstances of a hypothetical choice situation, but for the sake of argument suppose it does.

§2.3.3 Convergence and consensus

The contrast I have drawn between moral and prudential reasons is somewhat unfair. I have not quite been comparing like with like. To show how, I need to specify the examples a little more carefully. Consider again the case of Willow giving Xander a reason to refrain from stealing. Suppose, first, that Willow is a deontologist and Xander is a utilitarian. Willow tells Xander that he has a free-standing moral duty to refrain from stealing. Xander, however, believes that his only moral duty is to maximise overall utility; if stealing will maximise overall utility, he intends to steal. Compare this case with the case in which Willow gives Xander a prudential reason instead of a moral reason. Willow tells Xander that if he steals, she will shoot him, thus thwarting his goal of not being shot. On the basis of this comparison, it seems clear that the moral reason will be ineffective, but that the prudential reason will be effective.

However, in order to illustrate the point, I've rigged the game. Suppose, in the prudential reason case, that Willow had told Xander that his stealing of her property would thwart the achievement of one of *her* goals (continuing to have her stuff, for example). Clearly, although someone's goals are at stake, Willow's claim in this case does not give Xander a prudential reason to refrain from stealing. In order to provide a prudential reason for Xander, the claim Willow makes must refer to *his* goals. But surely something similar can be said in the moral reason case. Willow's mistake was not that she gave Xander a moral reason, but rather that she gave him a deontological moral reason rather than a utilitarian moral reason. Therefore, the lesson we should take from this discussion of politics and effective reasons is not that politics must be pursued through the use of prudential reasons, but that, in order to be effective, the moral reasons we offer to fellow citizens when engaged in politics must be reasons which they can affirm from within their own moral systems.

This sort of idea should be familiar – it resembles the Rawlsian idea of an overlapping

consensus.[38] Willow believes that she and Xander each have a free-standing moral duty to refrain from stealing, but she recognises that Xander is a utilitarian, and thus recognises that her deontological reasons will not be effective in motivating Xander to act as she prescribes. So, she offers him a utilitarian reason – an argument to the effect that refraining from stealing will maximise utility – even though she does not herself believe such a reason to be sound. Similarly, Xander thinks Willow's deontology is nonsense on stilts, but when interacting with her politically he can give her moral reasons which fit with her beliefs. The two agents do not agree on *fundamental* moral principles, but their moral systems overlap on the prohibition on stealing, and so there are moral reasons to refrain from stealing which can function as effective reasons for each of them.

While there is a resemblance, it is important to be clear that what I have just described is *not* what Rawls means by an 'overlapping consensus'. Rather, it is an example of what has come to be called a 'convergence' model of political agreement. The idea is that various moral systems can converge on the same prescription for quite different reasons. This is contrasted with a 'consensus' model, of which Rawls's is one.[39] In a consensus model, a prescription is supported by a reason which is shared by a number of moral systems. The prospects for an actual moral consensus among people in contemporary societies strike me as dim; and, of course, proponents of consensus models, such as Rawls, typically restrict the scope of their models to 'reasonable' people – they do not claim that there is going to be an actual consensus among all of the real people constituting an existing society. Convergence, on the other hand, looks more plausible, and convergence gives us a way for moral reasons to be effective in politics.

My argument has been that the reasons we use in politics must be effective reasons. To the extent that moral reasons are effective, I have no justification for excluding them. The argument for strict political contractarianism therefore rests on a certain

[38] Rawls, 2005: 133-172.
[39] See, for example, the discussion in Quong (2011: 263-264). Quong attributes this distinction to D'Agostino (1996: 30). Rawls (1997: 800) expressly rejects convergence. For a summary of the debate on convergence and consensus, see Vallier (2018: §2.3).

scepticism that it will be possible to find either a consensus or convergence of moral reasons sufficient to ground substantive prescriptions (as well as on a promissory note to the effect that sufficient prudential reasons can be found). That is to say, I doubt people like Willow and Xander will very often find that their moral reasons converge on the same prescriptions. Insofar as I am wrong about that – insofar as I overestimate the extent of moral disagreement, or the degree to which it is intractable – the argument I have presented will be unsound. And it is sometimes argued that, while there may be substantial moral disagreement about hard cases, there is striking *agreement* about some things (even quite fundamental things).[40] Is there really very much disagreement about the moral equality of persons, for instance?

There are two things to say in response to this. The first is that moral agreement is often more apparent than real. It is relatively easy to secure agreement for general principles such as freedom and equality, or basic rights, when one doesn't specify exactly how those principles will be interpreted, or what exactly their content is, or how they will be prioritised. Libertarians and egalitarian liberals agree on the moral equality of persons, but they have radically different understandings of what moral equality entails. The prospects for moral consensus or convergence look much worse if what has to be agreed is the full extension of a moral proposition.[41]

The second response is more conciliatory. If it really is the case that, in any population for which we are interested in generating political institutions, there is a consensus on some particular moral proposition (and on the full extension of that proposition), then I do not see any reason to exclude the proposition from political intercourse. In not actually being controversial, such a proposition would pass the bar I have set for political reasons – it would be perfectly capable of being an

[40] Dworkin (1996: 113) makes this point. Rawls (2005: 36-38) rejects this kind of claim.
[41] Quong, 2011: 29. Strictly, the full extension of a proposition might be an unreasonably demanding standard, given that *full* extension might have to cover contingencies that will never in fact occur. Nonetheless, I take it the point is clear: apparent agreement on the content of a proposition often conceals deep disagreements which arise in application. I thank Robert Jubb for helpful comments on this point.

effective reason.[42] Similarly, to the extent that a convergence model can generate effective moral reasons, such reasons can be the appropriate constituents of politics.

The convergence-consensus distinction can also be applied to prudential reasons. Genuine prudential consensus is likely to be uncommon, since it requires that I have goals which relate to the achievement of your goals.[43] But it is quite easy to think of circumstances in which our prudential reasons can converge – any commercial exchange qualifies. I will ultimately argue that certain political institutions are the outcome of a contractarian bargain between prospective co-citizens. Contractarian bargaining is itself a form of prudential convergence. In a successful bargain, the interests of the parties converge on the bargaining solution, which each party prefers to the pre-bargain *status quo*. Each party has a distinct prudential reason to accept the bargaining solution, but the various different prudential reasons the parties have converge on a single prescription for coordinated action.

There is no guarantee that convergence on a contractarian bargaining solution will be possible – it may turn out that we are not in the circumstances of justice after all.[44] The existence of such a solution is just as much a contingent fact of the world in which we find ourselves as the existence (or, as I have suggested, non-existence) of a viable convergence model for moral reasons. My background assumptions are that the prospects for prudentially-grounded convergence are pretty good, and that the prospects for morally-grounded convergence are pretty bad (and the prospects for morally-grounded consensus are virtually zero). Later on, I will offer some argument to the effect that the prospects for prudentially-grounded convergence are

[42] Compare Moehler, 2018: 19, 28, 163-165. Note one important way in which the standard I adopt here differs from Rawls's public reason test: Under Rawls's conception of public reason (2005: 212-254; see also *ibid*.: 39-40, 141, 142; and Rawls, 1997), it is sufficient to exclude a proposition from public reason that some possible reasonable co-citizen could reasonably not accept the proposition, even if none of one's co-citizens actually do reject the proposition. So, a population composed exclusively of orthodox Jews could not appeal to the Torah when engaged in political intercourse on Rawls's account of public reason, but could do so on my account (compare Quong, 2011: 151; and Larmore, 1990: 358 n3; see also Binmore, 1994: 146-147).

[43] Nonetheless, consensus is possible – consider a student who wants to do well in his exams, and his parent who wants him to achieve his goals. Compare the discussion of covariance in §7.6.

[44] §1.5 n37, §2.4.3, §3.2.1. See also the discussion of agents with pathological preferences in §5.5.2. My claim will ultimately be that prudential convergence is possible with most people, but probably not with all people. Some people, then, will be excluded from the social contract.

better than they are sometimes thought to be.[45] But beyond these considerations, and my brief remarks on the difficulties that arise for moral convergence accounts when moral beliefs are given their full extension, I will not argue for the background assumptions. As such, my argument should be understood as claiming that *if* the background assumptions hold, political contractarianism follows.

In circumstances of pervasive and persistent moral disagreement, it will often be the case that co-citizens will not share fundamental moral commitments. As such, what is a subjective moral reason for one may not be a subjective moral reason for another. It will often be the case, then, that moral reasons will fail to be effective reasons. By contrast, as long as we are in the circumstances of justice, it will be possible to give one another effective prudential reasons. Insofar as the moral reasons available to us are ineffective, and the prudential reasons available to us are effective, we have *i*-type prudential reasons when engaged in politics to give other agents *u*-type prudential reasons rather than *u*-type moral reasons. In short, we have *i*-type prudential reasons to be political contractarians.

The possibility of moral convergence offers an alternative to my approach. I am sceptical that moral convergence is possible to achieve in contemporary societies, but I have no objection to it in principle. The most straightforward way of demonstrating that my scepticism is unwarranted would be to present a case of moral convergence.

§2.4 Moral and prudential reasons revisited

The argument from effectiveness establishes that, given pervasive and persistent moral disagreement, citizens have *i*-type prudential reasons to give one another *u*-type prudential reasons when engaging in politics. That is, citizens have *i*-type prudential reasons to be political contractarians. Suppose, however, that we also

[45] §5.5.2, §6.6, §8.5.

have *i*-type moral reasons which apply to political action. This is the position taken by political contractualists. Contractualists argue that we (morally-)ought to justify our exercise of political power to those over whom we exercise it in terms which (if they are reasonable) they can accept.

Since I do not deny the existence of the *i*-type moral reasons to which contractualists appeal, the question at this point is whether those reasons defeat the *i*-type prudential reasons upon which political contractarianism relies. In this section, I examine and reject three arguments to the effect that they do.

§2.4.1 Self-directed reasons[46]

Suppose you are barbecuing one afternoon when your apron suddenly catches fire. The flames are spreading fast. Luckily, there is a pond next to you. I think your goal of avoiding serious burns gives you a compelling reason to jump into the pond.[47] I further think that this is a reason you have which is (all else equal) independent of any moral considerations. It is a prudential reason.

Suppose one doubts this. One might argue, instead, that the reason you have to jump is a moral reason. Perhaps the thought is that you have a moral reason to minimise suffering in the world whenever you can easily do so, and this includes your own suffering. So, what appeared at first to be a prudential reason is really a moral reason, albeit one directed at oneself rather than at others. This thought might then be generalised to all supposed prudential reasons. If pleasure (say) is good, and I have a moral reason to promote goodness, then I am acting in accordance with my moral reasons whenever I seek pleasure, or satisfaction, or any of the other things which typically motivate preference orderings. In this way, morality expands to occupy all available normative space. There is no room left for any prudential reasons upon

[46] I thank David Birks for helpful discussion of these points.
[47] Compare Korsgaard, 1996: 145-150.

which contractarianism can rely. Call this the *imperial* account of morality.[48]

The imperial account can be made more or less persuasive, depending on how it is developed. Suppose jumping would cause five other people to catch fire. Clearly, this would give you a moral reason not to jump. If one were then to say that you have *no* reason to jump, because the moral arithmetic has changed, I take it the imperial account would be absurd. Of course you have a reason to jump, whether or not that is what you (morally-)ought to do, all things considered. Nonetheless, your reason to jump need not be a prudential reason. Instead, one could say you have a *pro tanto* moral reason to jump, but that reason might be outweighed by other moral reasons you have (to avoid causing other people to catch fire).

There are two different understandings of the *pro tanto* version of the imperial account to consider. One might say that the advancement of *any* end an agent has counts as a *pro tanto* moral good. Alternatively, one might say that only the advancement of certain kinds of ends counts as a *pro tanto* moral good. Each option has some counterintuitive implications.

If the advancement of any end an agent has counts as a *pro tanto* moral good, and on that basis we say that the agent has a (*pro tanto*) moral reason to act so as to advance the end, then we are forced to say that agents have (*pro tanto*) moral reasons to do morally abhorrent things. After all, people can have genocide as an end. Whatever one thinks of the substance of this position, accepting it would make only a cosmetic difference to my argument. On this interpretation of the imperial account, what I have been calling prudential reasons would be a subset of moral reasons, the elements of which have special properties which distinguish them from 'other-regarding' moral reasons. With some substitution of vocabulary, my argument would proceed as before. [49]

[48] Though he would not express it in quite these terms, I take it this is (more or less) Raz's view (2011: 165-169).

[49] Moreover, consider the phenomenology of moral and prudential reasoning. If all normativity were moral – or if there were just undifferentiated reasons, as opposed to distinctly moral reasons and prudential reasons – why would the experience of identifying what I am calling one's prudential reasons differ so radically from the experience of thinking about moral goodness and duty? See the

If, on the other hand, one places perfectionist limits on what kinds of things preferred by agents count as moral goods, then further argument is needed. On this interpretation of the imperial account, the mere fact that an agent believes a thing to be good does not give him a reason to pursue it – one only has reasons to pursue things which are objectively good.[50] Plausible versions of this position will have it that the advancement of many ends will be *pro tanto* moral goods, because achieving these ends will be pleasurable, say, without also requiring acts which are intrinsically wrong or bad. Nonetheless, *some* ends will have to be excluded. Otherwise, there is nothing to distinguish this *pro tanto* version of the imperial account from the previous one.

In order to reject this view, I need cases in which an agent has a reason to act in a certain way, even though doing so is not even a *pro tanto* moral good. I think such cases can be found. Recall the distinction between an objective reason and a subjective reason.[51] My claim is that agents can have subjective reasons to act which do not track their objective reasons, and so do not issue in even a *pro tanto* moral good.

To illustrate, consider the case of Angela.[52] Angela is thirsty and she wants to drink something which will quench her thirst. There are two glasses in front of her. One contains a clear liquid. The other contains a dark liquid. Angela believes the clear liquid to be water, and the dark liquid to be oil. In fact, the clear liquid is petrol, and the dark liquid is cola. In this case, I claim Angela has a subjective reason to drink the clear liquid, even though doing so will not result in even a *pro tanto* moral good.

Suppose it is objected at this point that subjective reasons are beliefs about objective reasons; they are not themselves reasons. When we observe that Angela drinks the clear liquid, despite apparently being of sound mind, we can *explain* her action by

examples in the next section (§2.4.2), and particularly the discussion of supererogation.
[50] Raz, 2011: 165-169.
[51] §2.3.1, particularly n29.
[52] Lenman, 2009: §1; Williams, 1981: 102-103.

reference to her mistaken beliefs, but her action is not *justified*. At best, then, a subjective reason can be an explanatory reason. But a contractarian argument needs there to be non-moral reasons which are justifying reasons, not merely explanatory reasons.

This objection rests on an implausibly narrow understanding of justification.[53] It suggests that I can only be justified in doing what a perfectly informed agent would do. In particular, on such a view, we could not say that Angela is justified in complying with means-end rationality. Suppose, as before, that Angela (mistakenly) believes the clear liquid to be water, and the dark liquid to be oil. She (correctly) believes that drinking oil will make her feel worse rather than better. Suppose she nonetheless decides to drink the dark liquid, while still aiming to quench her thirst. We would not want to say that her action is justified simply because she was mistaken about the contents of the glasses, and through sheer luck managed to act in accordance with her objective reasons (by drinking what turns out to be cola). Or, at least, we want to be able to say that Angela's action was in *some* sense unjustified, because it represents a basic failure of rational deliberation.[54] Given Angela's beliefs, there is a (subjective) sense in which she is justified in drinking the clear liquid.[55] That is to say, she has a justifying reason to drink the clear liquid, even though no objective moral good can come of doing so.[56] Thus, the final version of the imperial account of morality can be rejected. Agents can have reasons to do things which are not moral reasons.

§2.4.2 Priority and redundancy

Suppose one thinks that the normativity of moral reasons is preemptive: Whenever

[53] Scanlon, 2003: 140; Quong, 2011: 141-142; Rawls, 1985: 229-230; Larmore, 2003: 376-377. In particular, note the distinction between justification *to* someone and justification *simpliciter*.
[54] Compare Raz, 2011: 141-142, 151, 155, 157.
[55] This does not, of course, imply that there is no sense in which she is justified in drinking the dark liquid. That Angela has a subjective justifying reason to drink the clear liquid is consistent with her having an objective justifying reason to drink the dark liquid.
[56] Compare Broome, 1999: 414-415; Broome, 2005a: 326-327, 333-335.

a moral reason conflicts with a prudential reason, the prudential reason must give way. If one also thinks that, in the political domain, it will always be the case that there is an *i*-type moral reason which applies to our circumstances, then on this view there will be no scope for *i*-type prudential reasons to determine what we (all-things-considered-)ought to do. The argument from effectiveness will be successful, inasmuch as it demonstrates that we have a (*pro tanto*) prudential reason to be political contractarians, but it will also be sterile. The reasons it offers us will be redundant – they will always be defeated.

It is sometimes taken to be a defining characteristic of moral reasons that they have preemptive force. The whole point of morality, on such an account, is to act as a constraint on our pursuit of our own interests.[57] So, suppose I am in a position to steal your wallet, can get away with doing so, and would benefit from the theft. I have a prudential reason to steal your wallet. However, I have a moral reason not to steal your wallet. What do I have most reason to do, all things considered?[58]

In this case, we would surely say that I (all-things-considered-)ought to refrain from stealing your wallet (absent some special story about my neediness, say). That is, in this case, my moral reason defeats my prudential reason. But consider a different case. Suppose you find yourself in a version of the trolley problem. A runaway trolley is hurtling towards a child. You can flip a switch to divert it down a siding. The catch is that your foot is stuck in the track which runs along the siding. If you do not flip the switch, the child will be killed. If you flip the switch, the child will be saved, but your leg will be crushed.[59]

I think you have a moral reason to flip the switch and save the child. All else equal, it seems clear to me that you have a prudential reason not to flip the switch. It is not

[57] Gauthier, 1986: 1, 3. Compare Rawls, 1999a: 3.
[58] I am assuming, for the sake of argument, that this is an intelligible question. It is not obvious that moral and prudential reasons are commensurable in the way I imply here. Even if they are not, however, given that we often have conflicting moral and prudential reasons, individuals themselves will have to find some way of adjudicating between them. Compare Neal, 1997: 195; Broome, 2005a: 324-325; Raz, 2011: 166. See also the discussion of incommensurability in §5.3, *infra*.
[59] Foot, 2002: 23; Thomson, 1986: 94-116; Singer, 1999.

obvious to me what you (all-things-considered-)ought to do. But, if I am right about the moral reason you have for flipping the switch (and *surely* you have a moral reason to save a child if you can), why isn't it obviously the case that that is what you (all-things-considered-)ought to do? If moral reasons always have preemptive force, surely it follows immediately that you (all-things-considered-)ought to flip the switch. If we don't think it immediately follows, then it seems that moral reasons do *not* always have preemptive force.

Assuming one is not inclined to bite the bullet, and say that you must flip the switch, there are two possible responses to this argument.[60] Either one could say that one has a countervailing moral reason not to flip the switch, and so this isn't a case in which a prudential reason trumps a moral reason, because it is actually the countervailing moral reason which is doing the work in defeating the moral reason you have to flip the switch. Or, one could say that one does not have a moral reason to flip the switch, and so there is no moral reason present which preempts the prudential reason you have to save your leg, and so this case is consistent with the claim that moral reasons (when they exist and apply) always have preemptive force.

Take the latter response first. It seems obvious that we do in fact have a moral reason to save the child, but perhaps we can make a 'specificationist' move at this point.[61] We would then say that the moral reason which applies to your situation has not been properly specified in the example I have given. The correct specification of the moral reason is as follows: you have a moral reason to save children in danger of being killed except when doing so would lead to one of your limbs being crushed. When we fully specify the moral reason in play, we find that there is a built-in exception which covers instances in which limbs are at stake, and so while it might appear that you have a moral reason to flip the switch, this is actually a mistake –

[60] If one is inclined to bite the bullet, we can simply increase the cost to you of flipping the switch. Unless one is willing to continue bullet biting all the way up to the point where flipping the switch will kill you, or condemn you to a lifetime of agony, the argument presented here will go through. There is nothing incoherent about such thoroughgoing agent-neutrality, but it does not strike me as a particularly plausible moral theory. In any event, I will address such cases at the end of §2.4.3. See also n67.

[61] Thomson, 1977; Shafer-Landau, 1995.

you don't.

Suppose we accept this line of argument. However, suppose you nevertheless decide to flip the switch, deliberately sacrificing your leg to save the child. If you have no moral reason to flip the switch, and no prudential reason to flip the switch, there appears to be no reason which could justify your decision to flip the switch. But that's implausible. Of course you have a moral reason to flip the switch and save the child. Doing so is not morally required – it is a supererogatory act – but duties are not the only moral reasons. The specificationist move, by stripping all moral reasons out of the scenario, leaves us unable to say that an act of heroism can be justified by the moral reason we have to save the child. This response can therefore be rejected.

That leaves us with the suggestion that you have a moral reason *not* to flip the switch which outweighs or defeats your moral reason to save the child. But what could that be? You might have a moral duty to avoid suffering (including your own), but how could any such moral reason outweigh the moral reason you have to flip the switch? Surely the child's life takes priority over your leg. If you doubt this, consider a slightly different scenario. In this case, there are two sets of tracks running parallel to one another. You are on one track and the child is on the other. Each track features a runaway trolley. As before, the child's life is at risk and your leg is at risk. I am a bystander, and I have time to save only one of you. Which do I have greater moral reason to save? Again, while it seems clear to me that I have (*pro tanto*) moral reasons to save both your leg and the child's life, it also seems clear that I have a greater moral reason to save the life of the child. If this is so, then when you are faced with the same choice, why wouldn't you have exactly the same set of moral reasons? Why would the fact that it's *your* leg give you a *moral* reason to refrain from flipping the switch which I wouldn't have? It is much less artificial to say that you have a prudential reason to save your leg which defeats your moral reason to save the child's life. Moreover, if the claim is that you have a moral reason to save your own leg which defeats your reason to save the child, where would that leave a heroism case? It cannot plausibly be maintained that you would be acting wrongly if you decided to save the child, but if the greater of the two moral reasons you have is to

save your leg, and if you act contrary to your greatest moral reason, it is hard to see how we can avoid the absurd conclusion that saving the child is morally wrong.

When discussing cases such as this one, we are often inclined to say that people have a moral duty to rescue others, but that the duty depends on two things: (i) the severity of the threat to the wellbeing of the person in danger; and (ii) the cost (or risk) to the rescuer of effecting the rescue. When the threat to the person in danger is relatively low, or the cost to the rescuer is relatively high, there is no duty to rescue. Potential rescuers, in such circumstances, have an 'agent-centred prerogative'.[62] They are (morally-)entitled to prioritise their own interests and projects over the wellbeing of the person in danger. They may permissibly fail to flip the switch (though, of course, they may also permissibly flip the switch – they don't have to exercise the prerogative).

This kind of account is popular, but does it give us a moral reason for action in the trolley case which outweighs our moral reason to save the child? Or is it, rather, an expression of the idea for which I have been arguing, the idea that moral reasons do not always have preemptive force?

I think we can probably tell the story either way. On the one hand, the prerogative ultimately arises from the rebellion of our intuitions against the conclusion that we always (all-things-considered-)ought to do things like flip the switch in the trolley case. Our intuitions, however, don't come with helpful labels attached indicating that they are intuitions about moral, rather than prudential, normativity.[63] And the 'prudentialised' understanding of the prerogative has the advantage of simplicity. On the other hand, we can understand the prerogative as a moral permission, one which is not itself subject to being overridden by prudential reasons. The first understanding of the prerogative implies that moral reasons do not always defeat prudential reasons. The second understanding allows us to retain the universally preemptive

[62] Scheffler, 1994 (and, in particular, *ibid*.: 20); Wellman, 1996: 214-216; compare Nozick, 1974: 34-35.
[63] Neal, 1997: 195.

quality of moral principles, but at the expense of building into them space for prudential reasons to actually determine our practical deliberations. Since moral reasons do not always defeat prudential reasons in practical deliberation, our question becomes whether or not they do so in the specific, substantive case of defining the bar of public reason. Our question is, does public reason fall within the scope of our agent-centred prerogative?

§2.4.3 Outcome-oriented and action-oriented reasons

The third objection to political contractarianism claims that, in the particular case of the reasons we have which apply to the definition of public reason (though not necessarily otherwise), our moral reasons defeat our prudential reasons. It may seem unlikely that I will be able to say anything general about this objection. People subscribe to different moral systems. Each moral system will have its own set of moral reasons which apply to the definition of public reason (or do not), and its own way of weighing moral and prudential reasons of various kinds. Since I cannot settle here the question of which moral system is correct, it may appear that all I can do is recommend a case by case approach, working out if the moral system to which this or that person subscribes holds that there are *i*-type moral reasons which defeat the *i*-type prudential reasons to which the contractarian argument appeals.

I think I can do somewhat better. I will argue that, for people who subscribe to a broad range of possible moral systems, there are reasons internal to those systems which decisively recommend political contractarianism. My argument will not show that everyone has decisive reasons to be a political contractarian, but it should cover the preponderance of cases we are actually likely to encounter. In concluding, I will comment briefly on the remaining cases.

To begin, suppose we divide moral reasons into two categories. *Outcome-oriented* moral reasons take the form of an injunction to bring about a certain state of affairs (or to bring it about that a certain state of affairs does not obtain). *Action-oriented*

moral reasons take the form of an injunction to do (or refrain from doing) some action.[64]

Insofar as moral reasons are outcome-oriented, the way they function in practical reason is exactly the way that other ends do.[65] The moral reason sets the end which the agent intends to pursue, and the agent then has a prudential reason to choose effective means to that end (because agents always have prudential reasons to choose effective means to their ends). This being the case, the argument from effectiveness applies. So, for example, suppose you believe you have a moral duty to ensure that people's basic rights against non-interference are respected. Your end is a state of affairs in which these basic rights are secure. Suppose, further, that you cannot accomplish this goal by yourself – you require the coordination of other people to set up a system of laws to achieve your aim. Thus, you have an *i*-type prudential reason to give other people effective reasons to coordinate with you in pursuing your end of securing basic rights. On the basis of the argument offered in §2.3, given the problem of moral disagreement, this means you have an *i*-type prudential reason to give *u*-type prudential reasons to prospective participants in your scheme. That is to say, given your outcome-oriented moral end, you have an *i*-type prudential reason to be a political contractarian. So long as the outcome which is the object of an outcome-oriented moral reason is one which requires the coordination of others, moral reasons of this kind will actually give rise to (instrumental) *i*-type prudential reasons in favour of political contractarianism.[66] This is true regardless of the content of the outcome-oriented moral reason. Any person who has *i*-type moral reasons which are outcome-oriented will have an *i*-type prudential reason to be a political contractarian, whether her moral commitments are Christian, utilitarian, or Nazi.[67]

---

[64] The 'outcome' and 'action' terminology is imperfect, but the distinction I am getting at is between actions an agent takes because they are instrumental to an outcome which has intrinsic moral value, and actions an agent takes because they have intrinsic moral value. See also Binmore, 1994: 310-312.

[65] Compare Rawls, 2005: 82-83.

[66] Compare Moehler, 2018: 19. In the event that the object of an outcome-oriented moral reason is one which does not require the coordination of others, it falls outside the scope of politics, and so outside the scope of political contractarianism.

[67] Compare McCabe, 2010: 156-158. This conclusion draws some of the sting from the bullet biting

So, if the *i*-type moral reasons we have which apply to the definition of public reason are outcome-oriented, the present objection can be dismissed, because the reasons in question do not weigh against the reasons upon which political contractarianism relies. However, not all moral reasons are outcome-oriented. Some are action-oriented. Suppose, for example, that you believe you have a moral duty to refrain from intentionally killing innocent people. The kind of moral reason generated by a principle like this is not so easily assimilated into the contractarian account. To be perfectly clear, the injunction I am considering is not, "Bring it about that no innocent person is intentionally killed." It is not even, "Bring it about that you never intentionally kill an innocent person." To the extent that the injunction can be translated in these ways, it is really relying on an outcome-oriented moral reason, and can be accommodated in the manner just described. The strictly action-oriented injunction with which I am concerned is, "Do not intentionally kill an innocent person."

In order to sustain the objection, an action-oriented moral reason has to do two things. First, it has to contradict the manner in which political contractarianism defines public reason; that is, it has to give us a different account of the *u*-type reasons which are admissible in politics. Second, it has to defeat the *i*-type prudential reasons which support political contractarianism.

It may appear that an action-oriented reason such as the injunction not to kill will struggle to meet the first of these conditions. After all, what does refraining from intentionally killing innocent people have to do with public reason? I can refrain from intentionally doing something all by myself – it isn't an action which requires a contribution from some other agent. As such, it is not a political action.[68] This is not something unique to the injunction not to kill. Many action-oriented moral reasons – including all of the classic deontological constraints against assault,

response I considered in §2.4.2 (in particular, n60), *supra*. Agent-neutral consequentialists can be led to political contractarianism by way of outcome-oriented moral reasons. So, it's only agent-neutral non-consequentialists for whom the bullet biting response implies a rejection of my view.

[68] §2.1.

torture, etc. – will be non-political for the same reason. If a moral reason is only governing my non-political actions, how can it contradict the manner in which political contractarianism tells me to act in the political domain?

Unfortunately, things aren't quite so straightforward. Recall that I have taken reason giving in politics to include both explicit speech acts and actions which intentionally provide incentives to other agents to behave in a certain way, and thereby amount to an instance of implicit reason giving. With that in mind, consider again the case of me pushing you off a bridge in order to communicate a political message to some audience.[69] Pushing you may be the most effective means I have to convince my audience to do as I say, and thus it may be the most effective *u*-type reason I can give. Nonetheless, it is the kind of action which could be ruled out by an action-oriented moral reason to refrain from intentionally killing innocent people. The action-oriented moral reason blocks me from engaging in a certain kind of implicit reason giving. In doing so, it contributes a partial specification of the content of public reason. Insofar as political contractarianism tells me to choose the most effective means to my ends when acting in the political domain, it can be contradicted by an action-oriented reason such as the injunction not to intentionally kill innocent people.

The fact that the duty not to kill potentially conflicts with the prescriptions of political contractarianism is not, in itself, sufficient to sustain the present redundancy objection. It must also be the case that the duty defeats the *i*-type prudential reasons which support political contractarianism. It is not obvious that this duty is absolute, and so it is possible for it to be overridden in favour of an *i*-type prudential reason, particularly if there is some important moral outcome at stake. Nonetheless, the action-oriented duty is sufficiently weighty that it could plausibly defeat our *i*-type prudential reasons in a significant range of cases.

The statement of the objection I have just given relies on the claim that, in order to

[69] §2.1.

contradict political contractarianism, it is sufficient for an action-oriented moral reason to block the *most* effective *u*-type prudential reason which political contractarianism recommends (for instance, pushing you off the bridge). But this isn't so. Political contractarianism, like any contractarian theory, requires that we traffic in prudential reasons.[70] So long as we (in-some-sense-)ought, all things considered, to give (only) *u*-type prudential reasons in politics, political contractarianism is satisfied. It does not require that we give the most effective *u*-type prudential reasons available.[71] Sustaining the redundancy objection requires blocking *any* *u*-type prudential reason which could defeat a competing moral reason, thus leaving us no option but to give moral reasons.

The force of the redundancy objection now depends on the answers to two questions. First, how likely are we to find ourselves in political circumstances where we have few prudential options, such that a weighty action-oriented moral reason with relatively limited scope, like the injunction not to kill, blocks all of the effective *u*-type prudential reasons which we could give? Second, do we have action-oriented moral reasons with a much more extensive scope, such that they could plausibly block all of the available prudential reasons, and if we do, do those action-oriented reasons have sufficient weight to defeat the *i*-type prudential reasons we have to be political contractarians?

Take the first of these questions, and consider action-oriented moral reasons like the classic deontological constraints: duties to refrain from killing, seriously injuring, torturing, enslaving – perhaps lying – and so on. With the possible exception of the prohibition on lying, these duties plausibly have sufficient weight to defeat an *i*-type prudential reason in favour of political contractarianism. It is certainly possible for us to find ourselves in circumstances where complying with all of these duties

[70] §1.4, and n26 in particular; compare §2.3.3 n42.

[71] It's true that *prudential reason* requires that one choose the most effective means, but I have not claimed that prudential reason is the only kind of reason. In the kinds of cases I am considering here, moral reason is limiting the range of options among which an agent may select on the basis of prudential reason. Political contractarianism requires that agents give one another only (*u*-type) prudential reasons in politics (again, as qualified by the considerations discussed in §2.3.3 n42), not that agents act only on the basis of (*i*-type) prudential reason.

renders us unable to give other people *u*-type prudential reasons of the kind political contractarianism demands. Still, so long as we stick with only the most central deontological constraints, the circumstances look quite unusual. In particular, they are an implausible characterisation of the contractarian choice situation. In order to block all *u*-type prudential reasons, the choice situation would have to be characterised in such a way that agents cannot make credible offers of positive assistance to one another. Such offers would be *u*-type prudential reasons, but would not be blocked by the classic deontological constraints, because the classic constraints are all negative duties. Even if we take some of the deontological constraints to be positive – to include a duty of rescue, say – it still seems implausible to think that agents in the choice situation will not be able to make offers above and beyond the level of assistance required by their positive duties.[72] Thus, while the central deontological constraints plausibly have sufficient weight to block political contractarianism, they are too narrow in scope to sustain the redundancy objection by themselves.

Consider, then, two examples of action-oriented moral reasons with a more extensive scope: a duty to refrain from threatening others, and something like Rawls's 'duty of civility'.[73] Given how central a role threats play as *u*-type prudential reasons, an action-oriented duty to refrain from threatening other people could have a very wide scope. In combination with the classic deontological constraints, such a duty could well rule out the bulk of the *u*-type prudential reasons an agent might have available in a contractarian choice situation. The agents might still have the option of making offers, but suppose we consider agents who lack the capacity to engage in productive exchange with others, and so cannot make credible offers.

Even so, this kind of reason cannot plausibly sustain the redundancy objection. It is not plausible to think that a moral duty we have to refrain from threatening other people defeats any of a range of our important outcome-oriented moral reasons.

[72] Wellman, 1996. This obviously doesn't hold if one is a thoroughly agent-neutral consequentialist. See n67, *supra*.
[73] 2005: 217-220. See also *ibid*.: 83-85.

Laws, for instance, are threats. That is to say, we typically have no problem threatening people when doing so is necessary in securing justice – consider again the case of a police officer threatening to shoot a person engaged in an unjust assault.[74] In fact, we are usually willing to go much further, and say that many of the classic deontological constraints do not apply if disregarding them is necessary in establishing a just state of affairs. We think it is (sometimes) permissible to kill in pursuit of justice.[75] While certain anarchists and pacifists might be willing to bite the bullet here, no liberal can do so.[76] Even if the duty to refrain from threatening people had the scope necessary to block political contractarianism, it would lack the required weight.

Unlike some of the other examples I have considered, the duty of civility has exactly the scope necessary to conflict with the *i*-type prudential reasons political contractarianism employs, because it is an action-oriented moral reason which specifically addresses the reasons we (morally-)ought to give one another when acting in the political domain.[77] The duty of civility is a duty to justify one's exercise of political power over one's fellow citizens to them, using (*u*-type moral) reasons which they could accept from within their own reasonable comprehensive doctrines.[78] In order to sustain the redundancy objection, the moral reason we have to act in accordance with the duty of civility would have to defeat the prudential reasons we have to be political contractarians.

As before, when set against our outcome-oriented moral reasons, this is implausible.

---

[74] §2.2.

[75] Torture and slavery are possible exceptions here. However, in order to sustain the redundancy objection, torture and slavery would have to be our only available prudential options, which is very unlikely.

[76] §4.1.

[77] This is true if we interpret the duty of civility as applying in non-ideal cases such as the real world. In those cases, the duty of civility will typically conflict with our *i*-type prudential reasons. If, on the other hand, the duty of civility is intended to apply only in a well-ordered society, there is no conflict between the duty and our *i*-type prudential reasons, because the fact of pluralism does not hold. See §1.5 n39; §2.3.3 n42; §3.2.2. See also Quong, 2011: 137-160. Moreover, note that a Rawlsian duty of civility will not apply to the whole of the political domain as I have defined it (Rawls, 2005: 135-136; Rawls, 1997; compare Quong, 2011: 273-287).

[78] Strictly, the duty is to be ready and able to give reasons of this kind (Rawls, 2005: 217). However, in the context of a hypothetical social contract theory, this amounts to the same thing.

If we have such reasons, they will weigh against the duty of civility. Given that outcome-oriented moral reasons might include such things as bringing it about that no one's basic rights are violated, we will need an overwhelming argument to show us that our duty of civility defeats our duty to bring it about that no one's basic rights are violated. But, intuitively, it isn't clear why we would think that a duty of civility could defeat any of a range of morally urgent outcome-oriented moral reasons.

In §2.4.2, I argued that, unless one is willing to insist on perfect agent neutrality, it is implausible to think that moral reasons always defeat prudential reasons in practical deliberations. Here, I've argued that it is no more plausible to think that moral reasons are always decisive in defining the bar of public reason. Both of these responses have relied on appeals to intuition, particularly about the existence of agent-centred prerogatives, and about the relative urgency of our outcome-oriented and action-oriented moral reasons. These appeals will no doubt be controversial. A thoroughly agent-neutral consequentialist might reject agent-centred prerogatives; a hardcore anarchist or pacifist might insist on absolute moral duties to refrain from making threats; an unrepentant contractualist might defend the priority (and real-world applicability) of the duty of civility.

There is nothing incoherent about these positions, and given what I have said about the persistence of moral disagreement, I cannot definitively refute them.[79] I can point to the intuitive cost of maintaining them, in terms of the prospects for pursuing our outcome-oriented moral reasons or the congruence of liberalism and agent-centred prerogatives. But, in the final analysis, I must concede that if a person believes herself to have a reason to give *u*-type moral reasons in politics – a reason which defeats all of her other reasons, prudential and moral – then nothing I say here can persuade her otherwise.[80] And so, as I said earlier, not everyone will have decisive (subjective) reasons to be a political contractarian.

This 'coverage' problem is not unique to political contractarianism. As I will discuss

[79] §2.3.2.
[80] Compare Rawls, 2005: 386-387.

in Chapter 3, political contractualists similarly exclude some proportion of citizens from their analysis, describing them as 'unreasonable'. People who adopt one of the moral absolutist positions I have described will function as contractarian analogues of contractualism's unreasonable citizens.[81] I discuss the impact they have on the contractarian theory in §5.5.2.

For a wide range of moral systems, the *i*-type prudential reasons we have to be political contractarians will be decisive reasons for action. They will not collapse into moral reasons, and they will not automatically be defeated by whatever *i*-type moral reasons applying to the definition of public reason those moral systems include. The most plausible and weighty of action-oriented moral reasons – the classic deontological constraints – are unlikely to conflict with the *i*-type prudential reasons we have to be political contractarians. And, perhaps most importantly, the greater the weight which our moral system attaches to outcome-oriented moral reasons whose pursuit requires coordination with other people, the greater the (*i*-type, prudential) reason we have to be political contractarians.

[81] Compare Moehler, 2018: 104.

Chapter 3

In this chapter, I contrast political contractarianism and political contractualism. There are two reasons for doing this. First, political contractualism remains the dominant form of social contract theory in contemporary political philosophy. In developing a contractarian alternative, I have borrowed much of political contractualism's structure. In §3.1, I give a precise description of the points of divergence between the two approaches. Highlighting these differences allows me to give a clearer statement of the contractarian theory, and of what is at stake between the two.

Second, in the course of setting out their own positions, contractualists have sometimes suggested that political contractarianism should be rejected. For the most part, their direct engagement with contractarianism has been very limited.[1] But, in §3.2, I present their concerns and argue that they do not – at this stage – give us grounds to reject contractarianism.

§3.1 Two conceptions of the problem of moral disagreement

Political contractarianism and political contractualism are both concerned with the problem of moral disagreement.[2] That is to say, neither theory sets out to identify definitive and comprehensive moral truth. Rather, each attempts to explain how people can justify particular terms of political coordination to one another, given that they differ in their fundamental moral commitments.

Rawls states his problem thus:

[1] The notable exception here is Barry (1989; 1995; see also §1.7, *supra*).
[2] Which is not to say that either theory is solely concerned with moral disagreement (§2.3.2; Rawls, 2005: 224). Nonetheless, each theory singles out moral disagreement for special attention.

> "[How] is it possible for there to exist over time a just and stable society of free and equal citizens, who remain profoundly divided by reasonable religious, philosophical, and moral doctrines?"[3]

This is a nuanced formulation. I want to pick out those elements which are most relevant for my purpose, and to set aside the others.[4]

First, it is important for Rawls that the political institutions which are the object of contractual agreement are just. Moreover, we are to understand this requirement in a substantive way – 'justice' is to be measured against our considered convictions in reflective equilibrium – rather than in a bare procedural way.[5] By contrast, a contractarian theory does not explicitly build in any such constraint on the output of its procedure. In Rawls's terms, it is asking how it is possible for there to be any kind of stable society, given that citizens will be profoundly divided on religious, philosophical, and (especially) moral questions.

This, however, is a little quick. My aim is to defend a contractarian account of *liberal* political institutions. I will argue that political contractarians should endorse institutions which are similar to those Rawls endorses. As such, I do think that political contractarianism provides us with an account of how there can be a political society which is both stable and just, despite the problem of moral disagreement (because, independently of anything I argue here, I think liberal political institutions are more or less just). But justice is not a requirement that is built into my methodology from the outset. Instead, I first propose to use a contractarian method to identify a set of political institutions – these are the institutions I believe can be intersubjectively justified given the problem of moral

[3] 2005: 4. I focus on Rawls's contractualism because it is the most prominent example in the contemporary literature, but the discussion will have more general application.
[4] In particular, we can identify a concern for: (i) persistence of institutions over time; (ii) justice; (iii) stability; (iv) the freedom and equality of citizens; (v) pluralism with respect to comprehensive doctrines; (vi) the persistence of said pluralism; and (vii) the reasonableness of said pluralism. I consider (ii), (iii), (iv), and (vii) in the text. (i) is not my present concern, although I will touch on it tangentially in §3.2.2. I take (v) for granted. Some contractarians – Moehler (2018: 1), for example – also accept (vi), but I will remain agnostic on it (§1.5 n39).
[5] Rawls, 2005: 8; Quong, 2011: 155.

disagreement. Only then do we consider whether or not we find the prescriptions of the theory an acceptable guide to action, all things considered. If it turns out that the only institutions which might be intersubjectively justified on contractarian grounds are institutions which are at such variance with our considered convictions that we cannot permit ourselves to act as the contractarian theory prescribes, then political coordination is impossible. That would be a tragedy, but I don't believe we will be faced with it, and Part II of my discussion is devoted to explaining why. So, this first difference between the contractarian and the contractualist theories should not be exaggerated. While justice is not a formal constraint in the contractarian theory, in most circumstances we will only be interested in the output of the contractarian procedure to the extent that we consider it sufficiently just.

A second element of Rawls's formulation can be dealt with in a similar way. Rawls requires that his theory provide the basis for a society of free and equal citizens. I have said that I understand 'liberal' political institutions to be political institutions which are in some sense shaped by the values of freedom, equality, and neutrality.[6] Just as I will show the contractarian argument to support just (because liberal) political institutions, I will show it to be consistent with a conception of citizens as free and equal. So, we can set aside this second distinction too.

This leaves two further elements of Rawls's formulation which reveal more significant differences between the contractarian and contractualist understandings of the problem of moral disagreement. The first is Rawls's conception of 'stability'. The second is his emphasis on addressing *reasonable* disagreement, rather than disagreement *per se*.

Rawls is seeking a theory which will show how a society characterised by disagreement can have political institutions which are 'stable' (and, of course, just). But he distinguishes between two different kinds of stability:

[6] §1.1, §4.1.

> "In one way we view stability as a purely practical matter: if a [political] conception [of justice] fails to be stable, it is futile to try to realize it. Perhaps we think there are two separate tasks: one is to work out a political conception that seems sound, or reasonable, at least to us; the other is to find ways to bring others who reject it also to share it; or failing that, to act in accordance with it, if need be prompted by penalties enforced by state power. As long as the means of persuasion or enforcement can be found, the conception is viewed as stable."[7]

Rawls rejects this 'practical' conception of stability. However, this is exactly the conception of stability I have in mind – stability understood as the practical problem of securing coordination within a set of political institutions, in response to which we are concerned to find ways of offering effective reasons to coordinate (including threats, should these be required) to people who do not share our moral beliefs.[8] Contractualism, on the other hand:

> "...is concerned with stability in a different way. Finding a stable conception is not simply a matter of avoiding futility. Rather, what counts is the kind of stability, the nature of the forces that secure it."[9]

---

[7] 2005: 142.

[8] Rawls, I take it, would consider this to be the "uninteresting Hobbesian" version of his question (*ibid*.: 391 n27).

[9] *Ibid*.: 142. There is an ambiguity in Rawls's text at this point. It is not clear if he means to say that contractualism is *not* concerned with stability of the first kind, and is *instead* concerned with stability of the second kind; or if he means to say that contractualism is not *solely* concerned with stability of the first kind, but is *also* concerned with stability of the second kind. Saying that stability is "not *simply* a matter of avoiding futility" (*ibid*., emphasis added) suggests, or at least leaves open, the latter reading. The remarks on the following pages (*ibid*.: 143-144, quoted *infra*) strongly suggest the former reading – that Rawls sees the two kinds of stability as strict alternatives. However, further on again (*ibid*.: 147, 148, 391-392), when Rawls contrasts the overlapping consensus with a *modus vivendi*, he seems to be at pains to insist that an overlapping consensus possesses stability of the first kind. This view is reinforced by the discussion of stability in *Justice as Fairness: A Restatement* (Rawls, 2001: 195). See also Rawls's definition of a well-ordered society (2005: 38-39), in which he makes it a necessary condition of a society being well-ordered that "unreasonable comprehensive doctrines... do not gain enough currency to undermine society's essential justice." This seems to be a worry about stability of the first (practical) kind, not the second kind. A similar point is made at *ibid*.: 391.

Rawls wants his political institutions to be 'stable in the right way'.[10] If the institutions are to be stable in the right way, support for them cannot be coerced – it must not follow from threats of sanctions (or, we might add, from offers which amount to bribes). Rather, each citizen must be able to support the institutions freely, because of reasons that arise from within her own set of moral beliefs:

> "The point, then, is that the problem of stability is not that of bringing others who reject a conception to share it, or to act in accordance with it, by workable sanctions if necessary, as if the task were to find ways to impose that conception once we are convinced it is sound. Rather, justice as fairness is not reasonable in the first place unless in a suitable way it can win its support by addressing each citizen's reason, as explained within its own framework. Only so is it an account of the legitimacy of political authority as opposed to an account of how those who hold political power can satisfy themselves, and not citizens generally, that they are acting properly."[11]

So, the contractualist and contractarian theories have different understandings of what 'stability' is, and thus of what the problem of moral disagreement is, and of what would constitute a satisfactory solution to it. Note that this difference cannot be set aside in the way that I have set aside as relatively unimportant the two earlier differences. The conception of liberalism with which I am working does not require that liberal political institutions be stable in Rawls's preferred, 'normative' sense. There are substantive differences between contractarianism and contractualism arising from their different understandings of stability.

---

[10] *Ibid*.: 38-40.

[11] *Ibid*.: 143-144. Quong (2011: 164) is more explicit. He understands "stable for the right reasons" to mean "endorsed by citizens for moral and not merely prudential reasons..." On Quong's view, what distinguishes an overlapping consensus of reasonable comprehensive doctrines from a "mere" *modus vivendi* is precisely that, in the case of the former, citizens will be able to endorse a political conception of justice on the basis of the moral reasons contained in their comprehensive doctrines. By extension, we can characterise a *modus vivendi* as a case in which the political conception of justice is endorsed on the basis of prudential reasons. Thus, the contrast between the contractualist and contractarian conceptions of stability is ultimately the contrast between an overlapping consensus and a *modus vivendi*. See also Rossi, 2010: 27; and §3.2.2, *infra*.

The second substantive difference is related to the role of reasonable disagreement. Recall that Rawls asks how it is possible for there to exist a just and stable society if citizens remain profoundly divided by *reasonable* doctrines. 'Reasonable' is here to be understood in a technical sense. For Rawls, an agent is reasonable if (and only if) she (i) is willing to propose fair terms of cooperation which it is reasonable to expect that other reasonable agents can accept, and is willing to abide by these terms if others also abide by them; (ii) recognises (what Rawls calls) the 'burdens of judgement'; and (iii) accepts the consequences of the burdens of judgement.[12] For the present purpose, we do not need to go into exactly what the various elements of this definition mean. It will be sufficient to note that, if an agent is reasonable, she believes herself to have certain moral reasons, and is committed to acting on them, assuming that certain conditions obtain (basically, that she is in a society with other reasonable agents who, by virtue of being reasonable, wish to cooperate with her in establishing political institutions). These moral reasons are (by definition) shared by all reasonable agents.

Political contractarianism is concerned with both reasonable disagreement and unreasonable disagreement (keeping in mind that 'unreasonable' is a technical term). What this distinction between the theories amounts to is a different way of characterising a 'constituency of justification' – the agents whose (unanimous, hypothetical) agreement is necessary to justify political institutions.[13] The contractarian requires that agents in the constituency are rational.[14] The contractualist requires that agents are both rational *and* reasonable.[15]

Combining the two substantive points of disagreement, we can see how contractualism approaches a solution to the problem of moral disagreement. First,

[12] Specifying exactly what Rawls takes to be the necessary and sufficient conditions of an agent being reasonable is not a straightforward task. In particular, the appearance of circularity occasioned by the presence of the term 'reasonable' in its own definition strikes me as unavoidable, given the various uses to which Rawls puts the term. I take it the conditions I have listed in the text are the core criteria, and will at any rate be an acceptable working definition. See Rawls, 2005: 48-61, 81-82.
[13] Chapter 1. I take the term 'constituency of justification' from Quong (2011: 2-8).
[14] This is not, however, all that it requires – see §3.2.1. For more on the rationality requirement, see §5.1 and §5.2.
[15] Rawls, 2005: 48-54, 81-82; Rawls, 2001: 191.

the contractualist is seeking political institutions which are stable in the right way – which each citizen can support freely, on the basis of reasons that arise from within her own set of moral beliefs. Second, the set of agents who count as citizens for the purposes of the stability criterion consists of all (and only) those agents who are (rational and) reasonable. Being reasonable, these agents (by definition) share certain moral reasons, even as they differ on other moral questions. If the moral reasons which they share support a particular set of political institutions, then those institutions will be stable in the right way – they will be supported by reasons which arise from within the sets of moral beliefs of the reasonable agents. And, of course, the problem of moral disagreement will then be solved, because the moral reasons available to the reasonable agents, simply in virtue of being reasonable agents, will coincide on a particular set of political institutions. On this crucial question, there will not in fact *be* any moral disagreement.[16] This coincidence Rawls describes as an "overlapping consensus of reasonable comprehensive doctrines" and it is the solution he offers to the problem of moral disagreement.[17] The final stage of the argument is then the demonstration that some particular set (or range) of political institutions does in fact follow from the moral reasons which reasonable agents (by definition) share.[18]

In contrast, the contractarian understands moral disagreement to be a practical problem. The task is to find ways to ensure that people coordinate on a particular set of political institutions, whether that coordination is mandated by their moral commitments or their prudential interests (if necessary, calculated in light of threats of sanctions). A set of political institutions is stable, on the contractarian account, to the extent that people do in fact have sufficient reason to coordinate.

---

[16] This slightly overstates Rawls's case. Rawls allows that there might be some disagreement on the details of a political conception of justice, but holds that all reasonable comprehensive doctrines will, in virtue of being reasonable, coincide on a 'family' of liberal political conceptions (2005: 164; also Rawls, 1997: 773-775; Enoch, 2005: 124; Rawls, 2005: 6-7, 140-141, 167-168; Quong, 2011: 162-163, 180-187). The basic point stands, however: The contractualist solution to the problem of moral disagreement is to say that such disagreement is not total or complete among *reasonable* citizens. (Because they share the relevant moral propositions, the moral reasons reasonable citizens give to one another will be effective reasons – see §2.3, *supra*.)

[17] 2005: 133-134.

[18] Given that the definition of reasonable agents is to some extent open, it can be adjusted (in reflective equilibrium) so as to ensure that some set (or range) of political institutions follows.

Given its practical orientation, the contractarian account has a wider scope than the contractualist account. It is not concerned only with that subset of agents who are reasonable. Rather, it is designed to address any rational agent with whom coordination is (possible and) required to achieve the end of establishing and maintaining political institutions. That is, the set of agents who count as citizens for the purposes of the contractarian account's stability criterion consists of both reasonable and unreasonable agents.[19] Assuming that the fact of pluralism means these agents will not share sufficient moral reasons to ground a set of political institutions, the contractarian solution to the problem of moral disagreement is to use normative materials other than moral reasons to generate convergence among citizens on a particular set of political institutions.[20]

In one sense, then, the contractarian conception of the problem of moral disagreement presents a less formidable challenge, since practical stability is significantly less demanding than normative stability. To put that another way, the challenge is less formidable because the contractarian account has a much wider range of tools (reasons) available to it. It is not limited to using reasons which are drawn from a narrow overlapping consensus of (reasonable) comprehensive moral doctrines. In another sense, though, the challenge is more formidable, since the contractarian sets out to give sufficient reason for coordination to all rational agents, whereas the contractualist account need only concern itself with supplying sufficient reason to all (rational and) reasonable agents.[21] The contractarian account has a greater array of tools, but it must also accomplish significantly more with them.

[19] Although not necessarily *all* such agents. As I have touched on briefly already (§2.4.3), and will discuss in greater detail later (§3.2.1, §5.5), there is an analogue in the contractarian account which corresponds to the unreasonable agents excluded from consideration in the contractualist account.
[20] §2.3.3.
[21] However, see n19, and the discussion of the prudential variant of the impossibility problem in §3.2.1.

§3.2 Contractualism *contra* contractarianism

Contractualist criticisms of a contractarian approach fall into two broad categories, roughly tracking the two substantive differences between their understandings of the problem of moral disagreement. First, there are objections based on the concern that, by failing to restrict itself to dealing only with *reasonable* disagreement, contractarianism will ground political institutions that are in some substantive sense defective or problematically contingent. Second, there are objections based on the idea that political contractarianism is a kind of *modus vivendi* theory, and so is 'political in the wrong way'.

§3.2.1 Constituency of justification objections

First among the objections is a concern about the kinds of institutions we will end up with if the constituency of justification for political institutions includes unreasonable agents. Consider Quong's view:

> "Political liberalism could require that political principles be accepted by each and every actual person who is bound by them, but this would cause serious problems. There are many ignorant, immoral, self-obsessed, or otherwise troublesome people in the world as we know it, and we do not want our political principles to be held hostage to their unreasonable demands."[22]

In political social contract theories, the justification of political institutions depends on the endorsement of the constituency of justification. Thus, selecting the

[22] 2011: 37; see also *ibid*.: 149-151. Note that the objections in this section which I draw from Quong are objections he offers against what he calls the 'external' conception of political liberalism (*ibid*.: 138-144). Quong does not explicitly offer these objections against a contractarian approach, but he does identify such an approach, describing it as a *modus vivendi* version of political liberalism (*ibid*.: 17). Given that the contractarian approach shares (a version of) the external conception's commitment to a relatively unrestricted constituency of justification, I take it the objections Quong presses against the external conception can be applied *a fortiori* to the contractarian approach.

appropriate constituency is a key step in the argument of a theory. The most straightforward constituency of justification would be the constituency of actual citizens subject to some particular set of political institutions. However, as Quong notes, defining the constituency of justification as the set of actual inhabitants of a society creates some serious difficulties.

Suppose there are immoral agents in a constituency of justification – for example, agents who deny the moral equality of persons – but we nonetheless need the assent of all members of the constituency to conclude that a set of political institutions is justified. On the plausible assumption that one of the premises upon which the justification of liberal political institutions depends is the moral equality of persons, it follows that if these immoral agents are included in the constituency of justification, it will not be possible to justify liberal political institutions to *all* members of the constituency of justification. In such circumstances, a social contract theory would be forced to reject liberal political institutions as unjustified. Given that immoral agents of various kinds do actually exist in contemporary societies, this is not merely a theoretical problem. It implies that the liberal political institutions of our own societies are not justified, and that even minimally decent political institutions would not be justified. If we reject this conclusion, but want to retain the social contract methodology, it seems we must restrict our constituency of justification so as to exclude the kinds of agents to whom things like universal basic rights cannot be justified. Contractualism aims to do precisely this by limiting its constituency of justification to all *reasonable* citizens.

There are actually three distinct problems here for a social contract theory which does not restrict its constituency of justification to something like 'all reasonable agents'. First, there is the problem of *insufficient justness*.[23] Suppose we have a constituency of justification in which some agents believe that brown-eyed people are intrinsically of greater worth than blue-eyed people. Given such a constituency of justification, it will not be possible to justify political institutions in which blue-

[23] Quong, 2011: 146, 151, 166-169; also Barry, 1995: 41-43, 45-46; Horton, 2010: 438; McCabe, 2010: 154-156; Sleat, 2013: 99-100; Vanderschraaf, 2019: 304-305, 310.

eyed people have the same rights, liberties, and entitlements as brown-eyed people. But substantive justice – as determined by our considered convictions in reflective equilibrium – demands that political institutions not discriminate between citizens on the basis of eye colour. Political institutions which do so discriminate are insufficiently just. To the extent that the contractarian approach does not restrict its constituency of justification in such a way as to ensure that it prescribes political institutions which are sufficiently just, it is said to be defective.

In the case of insufficient justness, the problem with the unrestricted constituency of justification was that the political institutions upon which all members of the constituency converged were substantively unjust. Perhaps even more likely, however, is the possibility that in the case of an unrestricted constituency of justification there will be no convergence – there will be no set of political institutions which can be justified to *all* agents in the constituency. This is the problem of *impossibility*.[24] To the extent that the unrestricted constituency of justification employed by the contractarian approach makes the justification of political institutions impossible, the approach is said to be defective.

Finally, suppose that despite an unrestricted constituency of justification, there is convergence on a set of political institutions (and so we are not faced with the problem of impossibility); and suppose further that, through rare good fortune, the political institutions which can be justified to the constituency are sufficiently just (and so we are not faced with the problem of insufficient justness). There is nonetheless a problem of *relativism*. The problem is that our political institutions are contingent on the beliefs of actual citizens in a way which we might suppose genuinely just political institutions would not be.[25] The contractarian approach is said to be defective to the extent that it is committed to claiming that the justification

[24] Quong, 2011: 139-140, 152-153, 189, 233-234, 314; also Horton, 2010: 443-444; Sleat, 2013: 101, 103, 105-106; Thrasher, 2019: 446-447.

[25] Quong, 2011: 151; compare Rawls, 2005: 23, 146-148, 391-392. See also McCabe, 2010: 134, 156-158; Rossi, 2010: 33-34. To make this objection, we need not assume that justice *fully* determines our political institutions. The objection is that, on the contractarian account, political institutions are more contingent than they should be (however much we take that to be). See n27, *infra*.

of political institutions depends on the endorsement of actual citizens, rather than on the conformity of the institutions with the true requirements of justice.

It should be clear how the contractualist's approach avoids each of these three problems. If the relevant constituency is all reasonable agents, then the agents have sufficient commitments in common (by definition) to ground an overlapping consensus of reasonable comprehensive doctrines, and that consensus issues in determinate prescriptions for political institutions.[26] This avoids the problem of impossibility. Given the nature of the commitments which the reasonable agents share, we will also avoid the problem of insufficient justness. The basic commitments which define reasonable agents are chosen so that they will cohere with our (or, at least, with Rawls's) considered convictions about substantive justice. Finally, insofar as the constituency of justification in the contractualist model is understood to be a hypothetical constituency encompassing all possible reasonable agents, we will avoid the problem of relativism. Justification by appeal to the constituency of reasonable agents is simply a device for helping us to tease out the substantive content of just political institutions – our political principles are never dependent on the beliefs of any actual people, and therefore are not problematically contingent.

What should a contractarian make of these three concerns? In §3.1, I have already sketched a reply to the problem of insufficient justness. I agree that *if* the political institutions ultimately recommended by the contractarian theory for typical circumstances are insufficiently just, that would provide grounds for questioning the theory. However, my argument will be that the contractarian theory recommends liberal political institutions for typical circumstances, and liberal political institutions are sufficiently just. The problem of insufficient justness is a sensible concern to have about the contractarian project, but my aim in Part II will be to show that the concern is ultimately unfounded.

[26] Rawls, 2005: 48-61, 81-82. See also n18, *supra*.

Now, this way of responding to the problem of insufficient justness in fact makes the relativism problem all the sharper. My claim will be that for a range of cases and, most importantly, in the circumstances in which citizens of contemporary societies find themselves, *it will turn out that* the contractarian approach recommends liberal political institutions. But, without question, this argument concedes that the prescriptions of the contractarian approach are contingent on the beliefs of actual citizens. It will turn out, I argue, that we (prudentially-)ought to have liberal, and hence just, political institutions, but it could have turned out otherwise were circumstances sufficiently different. Is this contingency something which should worry a contractarian?

Well, no. It would be problematic to argue – as a moral relativist might – that the political institutions we (morally-)ought to endorse depend for their justification (in any deep way) on what certain fallible moral agents happen to believe.[27] But it would be problematic precisely to the extent that we think it unlikely that *moral* truths are contingent in such a way. Political contractarianism is not looking for moral truths, however. Rather, it is telling us what we (prudentially-)ought to do. The prescriptions of prudential reason are, by their nature, contingent on the preferences of the agents engaged in it. The objection from relativism is misplaced, because relativism is a worry we have specifically about moral theories being contingent in particular ways. No similar concern attends theories which trade in prudential reasons.[28]

Of the three, this leaves the problem of impossibility. Given that contractarian justification will depend on the preferences of actual citizens, it may turn out that there is no set of political institutions which can be justified by the lights of each and

[27] By contrast, 'shallow' dependence seems unproblematic (Thrasher, 2019: 444, 447). For example, it doesn't seem that principles of justice determine whether elections should be held every four years or every five. Such a decision is properly within the discretion of a democratic citizenry, and so the precise political institutions governing their society will, to that extent, depend on the beliefs of some actual citizens. The dependence here is shallow because the principles governing the application of the discretion are not themselves dependent on the beliefs of actual citizens – the deeper principles which structure the society are not dependent, and thus the contingency is not problematic.

[28] I have assumed here, without argument, that moral relativism is an implausible metaethical position; but I take it that anyone who thinks moral relativism is plausible could not press an objection from relativism against me.

every person in the society.

There are two variants of the impossibility problem. On the *moral variant*, the problem is that citizens have different, and conflicting, foundational moral commitments. Intersubjective justification necessarily proceeds from shared premises, and reasonable and unreasonable citizens do not share the necessary premises to morally justify political institutions to one another.[29] Given the persistence of moral disagreement, it is impossible to morally justify any particular set of political institutions to all citizens in a society.

This variant of the impossibility problem will be familiar – I made essentially the same point in §2.3. There, it served as a component of my argument in favour of using prudential reasons rather than moral reasons in politics. Obviously, then, this variant of the impossibility problem is no threat to the contractarian project. That project is *motivated* by the impossibility of moral justification in circumstances of pervasive and persistent moral disagreement; and what it is motivated to do is to appeal to prudential justification instead of moral justification. On this understanding of the complaint from impossibility, the complaint actually complements political contractarianism by illustrating the necessity of appealing to prudential reasons.

There is a second variant of the impossibility problem – the *prudential variant* – which cannot be dismissed so readily. As I have noted, the prescriptions of a contractarian theory are contingent on the preferences of the agents involved. Political contractarianism prescribes a social contract which is essentially a rationally-mandated compromise between the various preference orderings held by the parties to the contract.[30] This compromise is selected by means of a bargaining theory.[31] The bargaining theory operates on the assumption that coordination

[29] Quong, 2011: 141-142, 181; Rawls, 2005: 100-101; Vallier, 2019: 114-115.
[30] Including outcome-oriented moral commitments, understood for the purposes of the contractarian procedure as elements in the parties' preference orderings. See §2.4.3, §5.1. Compare McCabe, 2010: 133.
[31] Chapter 6.

between the parties to the contract will yield each of them a more preferred state of the world than they could achieve without coordination. That is, I assume coordination is *productive*. The prospect of receiving gains from coordination is what gives each agent a prudential reason to coordinate.

However, it is no more than an assumption that coordination between the parties is productive. The preferences of the parties need not be structured in such a way as to make coordination productive, and if they are not, then there will be no gains for the bargaining theory to identify and distribute. In such circumstances, there are no political institutions which can be the object of simultaneous prudential justification to all agents in the constituency of justification.

There are three cases to consider here. The easiest case is one in which the productive coordination assumption holds for all of the agents we are discussing. In such a case, there is no impossibility problem. There are then two other feasible cases. Either none of the agents can coordinate productively, or some but not all of the agents can coordinate productively. The case in which none of the agents can coordinate productively is highly unusual. Consider what it would mean for me to be completely incapable of affecting the realisation of *any* of your ends in *any* way, and *vice versa*; now imagine this is true for *everyone*. We can construct such scenarios, but they are not a plausible characterisation of political society.

The more interesting case is when productive coordination is possible between some but not all agents. Recall that a constituency of justification is that set of agents whose (unanimous, hypothetical) agreement is necessary for political institutions to be justified on the basis of a social contract theory. In the present case, the impossibility problem is that, due to a limited scope for productive coordination, not all members of the constituency of justification can be given sufficient (prudential) reason to agree to a particular set of political institutions. If a contractarian theory cannot secure the agreement of all members of its constituency of justification, it fails on its own terms. While it is implausible to claim that *no* productive coordination would be possible in a society, it is much more likely that a society will contain at least some people

among whom productive coordination is not possible.

On the basis of the discussion in §3.1, it might be thought that the constituency of justification for a contractarian theory is the set of all rational agents. But, strictly, this isn't so. Rather, the constituency of justification is the set of all rational agents *among whom productive coordination is possible*. Only with its constituency of justification so defined can the contractarian theory avoid attempting to do the impossible. But, with the constituency of justification so defined, it will always be possible to give all agents of the constituency prudential reasons to endorse some set of political institutions. Thus, there cannot be a prudential impossibility problem.[32]

Of the three constituency of justification objections to political contractarianism, neither the problem of relativism nor the problem of impossibility (in either variant) is persuasive. An objection from insufficient justness remains and answering it will be the task of Part II.

§3.2.2 Political-in-the-wrong-way objections

The second set of objections to political contractarianism are somewhat disparate. For ease of exposition, I organise them around the thought that there is something wrong with the idea of a *modus vivendi*, and that since political contractarianism is an example of a *modus vivendi* theory, there is therefore something wrong with political contractarianism.

'*Modus vivendi*' means, literally, a 'mode of living', but it carries the implication of adversaries finding a way to live with one another, at least temporarily – a way of peacefully coexisting despite the potential for conflict between them. In the context

[32] Compare Moehler, 2018: 101-102, 104. See n19, *supra*. This way of drawing the boundary of 'unreasonable' has the advantage of being much more precise than Vallier's 'deflationary' approach to reasonableness, for example, though of course he is putting the concept to a different use (2016: 609-610). For further discussion, see §5.5.2, *infra*.

of social contract theory, it signifies a state of affairs in which various factions support a society's political institutions as a second-best compromise, simply because each faction lacks the power to impose its most preferred state of affairs by force.[33]

We might think this is problematic for two reasons. On the one hand, if our justification for a particular set of political institutions is contingent on the balance of power among competing factions in a society, then should that balance change, the institutions justified by our theory could also change. Insofar as the balance of power could change rapidly and frequently, this would lead to the society's political institutions being profoundly unstable – not in any moralised Rawlsian sense, but in the perfectly mundane, practical sense of stability with which political contractarianism is supposed to be concerned. If political contractarianism were to imply that the political institutions justified in a society were changing from day to day, the theory would be defective by its own standard of stability.[34] On the other hand, even if this problem of *practical stability* did not arise, a *modus vivendi* would imply political institutions justified by nothing more than a calculation of self-interest, or worse, a threat of violence. This, we might think, is the wrong kind of reason to have when deliberating about politics. We can call this the problem of *normative stability*.

One point which I should concede at the outset is that political contractarianism is in fact a form of *modus vivendi* theory. It is expressly committed to the claim that political institutions represent a compromise between citizens who have conflicting and incompatible political visions but need to coordinate with one another to pursue their ends.[35] So, concerns about *modus vivendi* accounts are certainly relevant.

Consider first the problem of normative stability. For Rawls, it is important that a normative political theory not be 'political in the wrong way'.[36] How, he asks, can

[33] Neal, 1997: 185-201; Rawls, 2001: 192-193; McCabe, 2010: 126, 133-134; Rossi, 2010: 21, 26-28; Sleat, 2013: 90, 94-98; Moehler, 2018: 176-179; Vallier, 2019: 107.
[34] Compare Rossi, 2010: 27, 33-34.
[35] *Ibid.*: 21; Quong, 2011: 17.
[36] 2005: 40, 142.

a liberal political regime gain sufficient support from its citizens to be stable? One way of securing this support would be to "look to the comprehensive doctrines that in fact exist [in the society] and then draw up a political conception that strikes some kind of balance of forces between them."[37] This is what political contractarianism does. By designing one's political conception so that it appeals to the most prominent factions in one's society, one can ensure that the regime regulated by the conception will have the support it needs to be (practically) stable. But this would make the political conception 'political in the wrong way', by which we mean a vulgar political compromise – a *modus vivendi*.[38]

This much tells us that being a *modus vivendi* makes an account political in the wrong way, and therefore defective, on Rawls's account. It doesn't tell us why Rawls thinks that being a *modus vivendi* makes an account defective. What exactly is *wrong* about the way in which a *modus vivendi* is political?

Getting at the answer to this question requires that we go back to the starting points of political liberalism. Rawls tells us that "the fundamental organising idea of justice as fairness... is that of society as a fair system of cooperation... Fair terms of cooperation specify an idea of reciprocity... [and] the idea of reciprocity is not the idea of mutual advantage."[39] Fair terms of cooperation are to be understood as the object of a contractual agreement – this is a contractualist theory, after all – but the agreement is not simply a matter of bargaining for mutual advantage. Rather:

> "[The] agreement, like any other valid agreement, must be entered into under the appropriate conditions. In particular, these conditions must situate free and equal persons fairly and must not allow some persons greater bargaining advantages than others. Further, such things as threats of force and coercion, deception and fraud must be excluded."[40]

[37] *Ibid*.: 39.
[38] *Ibid*.: 39-40, 142; Barry, 1995: 41; Neal, 1997: 185-187; McCabe, 2010: 133, 156-158; Vallier, 2019: 107-110.
[39] 2005: 15, 16, 17.
[40] *Ibid*.: 23.

A *modus vivendi* account is political in the wrong way because it does not necessarily represent an agreement entered into under the appropriate conditions. Any such agreement is (at least potentially) *invalid* – it lacks the prescriptive force that agreement is typically taken to have.[41] It lacks that force because it is based on threats, or on deception, or because some parties to the agreement have more bargaining power than others, and that is unfair. The kind of practical stability with which *modus vivendi* accounts are concerned is secured by force or fraud or exploitation of unfair bargaining positions. By contrast, the stability with which Rawls is concerned is not simply a practical matter: "what counts is the kind of stability, the nature of the forces that secure it."[42] The nature of the forces which secure a *modus vivendi* – force, fraud, and exploitation – makes it impossible for such an arrangement to be legitimate. Insofar as what we want from our theory is an account of the moral legitimacy of political institutions, *modus vivendi* accounts can't do the job.[43]

The response to this objection is straightforward. Political contractarianism is not intended to give an account of the moral legitimacy of political institutions. If we understand Rawls to be trying to work out what justice demands of us in (what Rawls takes to be) the ideal circumstances of a well-ordered society, then the contractualist procedure is plausibly a good way of getting at an answer to that question.[44] Nevertheless, the contractarian theory is interested in a different question. It is asking what we (all-things-considered-)ought to do in the non-ideal circumstances of the societies in which we actually find ourselves.[45] On this understanding, the problem

---

[41] §1.2.
[42] 2005: 142.
[43] *Ibid*.: 143-144. Compare the discussion of being 'political in the wrong way' in *Justice as Fairness: A Restatement* (Rawls, 2001: 188, 191). This line of argument in Rawls can be traced through discussions in *A Theory of Justice* (1999a: 10-15, 112-118, 122) and "The Basic Structure as Subject" (2005: 265-269). See also Rossi, 2010: 30-36.
[44] Rawls, 2001: 8-9.
[45] If it turns out that we are in a well-ordered society, political contractarianism's basic assumption about the existence of pluralism doesn't hold – in such a society, many more moral reasons can be effective reasons, and so political contractarianism will be consistent with political contractualism. On my view, political contractualism is applicable in the special case of a well-ordered society, and political contractarianism is a more general theory which can cover both ideal and non-ideal circumstances.

of normative stability is a version of the problem of relativism.[46] There would be a problem if the contractarian theory were claiming that morally ideal political institutions are dependent on the balance of power in a society. But it isn't.

Now, if there were some kind of immediate connection between what justice demands of us in ideal circumstances and what we (all-things-considered-)ought to do in non-ideal circumstances, this response would be too quick. However, no such connection exists.[47] What Rawls's own theory tells us about non-ideal circumstances is that we have a natural duty to "assist in the establishment of just arrangements when they do not exist, at least when this can be done with little cost to ourselves."[48] On its face, this is a remarkably (perhaps implausibly) permissive understanding of the moral reasons we have when we find ourselves in non-ideal circumstances – the 'little cost' provision seems to leave open the possibility that (in a contractarian procedure, for instance) we may permissibly prioritise the advancement of many of our non-moral ends over the promotion of our conception of justice. But even on a more robust reading, the natural duty to assist in the establishment of just institutions when we are in non-ideal circumstances looks like it is perfectly compatible with the account I have given of political contractarianism, because it is exactly the kind of weighty, outcome-oriented moral reason which I have said political contractarianism can accommodate. Thus, we can set aside the problem of normative stability.

The other concern about *modus vivendi* accounts is the problem of practical stability. Recall that this is a concern about the way in which the content of justified political institutions might change frequently or in an unreliable way as it tracks shifts in the balance of power in a society. Although it is not his central complaint, Rawls articulates something like this objection in a number of places.[49] A

[46] §3.2.1.
[47] Enoch, 2015: 125.
[48] 1999a: 293-294.
[49] 2005: 148-149, 392. See also Rawls, 2001: 115, 192-193, 195; Larmore, 1990: 346; Barry, 1995: 37-39, 41-42 (note b); Neal, 1997: 194-195, 198-199; Horton, 2010: 440-441. This worry has a classical pedigree – see Aristotle's discussion of 'disproportionate increase' in Book V, Chapter (iii) of *The Politics* (1981: 302-303).

contractarian might respond by saying that (absent rare events like wars) the balance of power between advocates of different moral traditions in a society is unlikely to change quickly; rather, it will change gradually, in response to the evolving demographics of the society from one generation to the next. And is it really such a bad thing that it does? Perhaps our political institutions (in-some-sense-)ought to reflect the evolution of our societies.[50]

This response is unsatisfying, though, for two reasons. It's true that the relative strength of different groups in a society is unlikely to shift rapidly, and even less likely to rapidly shift often. However, in a society that is sufficiently fragmented there will be no faction which can exercise hegemonic power, and in the absence of a hegemonic group, the state (if there is one) will be controlled by a coalition of factions which can in combination exercise sufficient power to determine the political institutions of the society despite the opposition of groups outside the dominant coalition. While the underlying demographics of a society don't often shift rapidly, *coalitions* between groups in the society *could* change from one day to the next, and could do so frequently, since coalitions depend simply on how the various factions decide to combine for the purpose of exercising political power. Given the assumption of widespread pluralism with which the contractarian theory begins, it is not in a position to rule out this kind of practical instability.

The second reason to be sceptical of this response is that it creates troubling incentives for groups which do manage to achieve control of the state. In order to remain in control, dominant groups will have to ensure that the ideological composition of the society will always favour them, and the most obvious way to do so is by eliminating dissent – perhaps by means of cultural imperialism, totalitarian perfectionism, ethnic cleansing, or genocide.[51] Obviously, a contractarian theory which recommended this kind of action could not be a liberal theory. As such, I will give a different response. However, that response depends on elements of the contractarian theory I have not yet introduced, so I will defer further discussion of

[50] This is the main idea underpinning Muldoon's account (2016; see §1.6.4, *supra*).
[51] Margalit, 2009: 194-197; Moehler, 2018: 86.

practical stability.[52]

I have now examined five objections to political contractarianism. Three of these – the problem of impossibility, the problem of relativism, and the problem of normative stability – can be set aside. Two concerns remain outstanding: the problem of insufficient justness and the problem of practical stability. These problems define the parameters of the argument I will present in Part II. The contractarian theory developed there should issue in prescriptions for political institutions which are practically stable and sufficiently just.

[52] §8.4.

Part II

Justice as Mutual Advantage

Interface

Citizens in contemporary societies are committed to a wide variety of ends, including various conceptions of the good and of justice. This is the fact of pluralism. Pursuing these ends is the highest of a citizen's priorities. Nonetheless, a citizen typically lacks the power to achieve her ends in the absence of a willingness, on the part of others, to facilitate her in doing so. She typically needs either the positive assistance of others, or for others to refrain from interfering with her actions.

Given that citizens need one another's assistance or non-interference to pursue their ends, it will often be possible for them to coordinate their actions in a way which leaves each better off. This is the fact of moderate scarcity.[1] In the simplest cases, coordination of this kind might amount to a principle of non-interference: I refrain from interfering with you, if you, in return, refrain from interfering with me. It might also take more complicated forms: I gather coconuts for you in the afternoon, if you sing for me in the evening; I recognise a certain pile of coconuts as your 'property', if you chase off any wolves which threaten me; I use my medical talents to keep you healthy, if you refrain from using your martial talents to render me unhealthy.[2] Such coordination can ultimately extend to highly complex social, economic, and political institutions.

Assuming there is the possibility of productive coordination between us, it is likely that it can take many forms. So, for example, I could gather five coconuts for you, and you could sing for an hour. Or, I could gather ten coconuts, and you could sing for ten minutes. Suppose that either of these options is better for both of us than a situation in which there is no gathering and no singing. Nonetheless, I will prefer the former pattern of coordination, and you will prefer the latter (assuming I like more singing and less gathering, and you prefer the reverse). Thus, there is a bargain to be struck between us, specifying some rate of exchange of gathering for singing.

[1] §5.5.1.
[2] Moehler, 2018: 169-170.

More generally, suppose we are a group of agents. The world in which we live is marked by a number of features. Among these are the fact of pluralism, the fact of moderate scarcity, and the absence of any coordination. Given these facts, it is possible for all of us to benefit if we move from a world without coordination to a world with it. Let's further suppose that we are pretty ordinary people. We have quite typical ends. No-one likes being assaulted, and everyone needs secure control of certain resources in order to advance her ends. There is, again, a bargain to be struck between us. The terms of that bargain will probably include rules governing the use of force against one another; a regime regulating material resources; and a system of enforcement to which people make a specified contribution. That is, the terms of our bargain will specify that certain people have certain legal rights, there will be some kind of property system, there will be a system of law, and a system of taxation to provide for its enforcement.[3]

At the most basic level, this is the picture of political society which a contractarian theory offers us. Political society is understood as a joint venture between a collection of agents, each of whom enters into the venture for the purpose of advancing whatever ends she happens to have. The claim is not, of course, that we have ever actually sat around and discussed a bargain. Rather, the claim is that a particular form of political society is justified if it is a form to which each of us could have agreed, had we ever so sat down, well-informed and sound of mind, to settle the terms of our coordination.

More precisely, a contractarian theory tells us that political society of a particular form is justified if it would be chosen, from a range of options, in a suitably specified choice situation, by an idealised version of each real citizen. Presenting such a theory requires that we develop an understanding of several things: how to characterise an idealised version of a citizen, how to characterise the situation in which the hypothetical choice is to be made, and how an idealised agent would

[3] I assume a statist solution here. I discuss anarchism in §5.5.

respond to being in such a situation. By putting these things together in a coherent way we construct a *model* – a representation of the world designed to help us think systematically about how the world works and what we (in-some-sense-)ought to do.[4] My purpose in Part II is to defend a contractarian model which will tell us what terms of coordination a rational social contract would contain.

Before going any further, some brief remarks on the use of models in normative political philosophy are in order. A model is a simplified version of the world. By simplifying the world, a model makes it easier for us to understand. A good model – a useful model – accurately captures the relevant features of the world and excludes irrelevant features. The more we can exclude, the simpler the model. However, if in pursuit of simplicity we exclude relevant features of the world, the model is to that extent flawed. Building a model thus involves two questions of judgement. First, we need to decide which features of the world are relevant; second, we need to decide on some tradeoff between the complexity of the model and its accuracy.

There are two different kinds of models to consider here.[5] *Descriptive* models are intended to give us an understanding of how the world works. They describe the world and explain it to us. *Prescriptive* models are intended to tell us how we (in-some-sense-)ought to act. Normative political philosophy is concerned with prescriptive models.

Inaccuracy – failure to represent the world as it actually is – is always a (*pro tanto*) defect in a descriptive model. However, this isn't true of a prescriptive model. Suppose I am playing poker, and I am trying to decide whether to call or fold on a particular hand. A model could help me make the decision by recommending an optimal strategy, given my circumstances. However, what I want the model to do is tell me what an idealised poker player would do in my circumstances. Suppose, instead, the model were to tell me what I, with all of my irrational biases, would do in those circumstances. This model would provide a perfectly accurate description

[4] Vallier, 2019: 98-99.
[5] Compare Binmore, 1994: 334-335; Moehler, 2018: 96, 111.

of the world. As a prescriptive model, it would be entirely useless. The whole point of building a prescriptive model is to idealise in certain respects – I want to know what someone who is better at poker than I am would do, so that I can copy her. Thus, it is not necessarily a defect in a prescriptive model that it does not accurately reflect the world.[6]

The contractarian model I will develop is a prescriptive model. It is far from a perfectly accurate representation of the real world. Nonetheless, some of its inaccuracies are not defects. I am not perfectly rational, but what I want to know is what a perfectly rational agent would do when choosing the terms of political society, because a perfectly rational agent is, by definition, an agent who acts so as to maximally advance his or her goals. This is what the contractarian model tells me – what I would do if I were the kind of agent who always acts so as to maximally advance his goals. However, other inaccuracies in the model need some defence. For instance, it is unproblematic for the model to represent *me* as perfectly rational (or so I will argue). But, if it represents everyone around me as perfectly rational, when they are not, it is to that extent defective. When playing poker, the optimal response to a rational opponent and the optimal response to an irrational opponent will not always be the same thing.[7] Inaccuracies of this latter kind must be defended either as worthwhile simplifications, as reasonable presumptions in conditions where we lack full information, or as irrelevant to the features of the world the model is designed to capture.[8]

There is an ongoing debate among contemporary theorists about the manner in which it is most appropriate to idealise when constructing hypothetical social contract theories.[9] I want to pick out three strands of this discussion for

[6] Compare Enoch, 2005: 781 n47; Enoch, 2015: 118-120; Thoma, 2015: 3338; Moehler, 2018: 31-32. See also §7.3 and §7.4, *infra*.
[7] Binmore, 1994: 211 n34. For a much more general worry of this kind, see Broome (2005a: 333-335).
[8] Enoch, 2005; Enoch, 2015: 124-125.
[9] Among many others, see Enoch (2005, 2015), Gaus (2011: 232-258, 263-333; 2014), Quong (2011: 137-160; 2014), Vallier (2014: 145-180; 2016; 2019: 97-103), Thrasher (2019), and Billingham and Taylor (MS). For quick but helpful overviews, and further references, see Quong (2017: §3) and Vallier (2018: §2.4). Much of the debate has been occurring in the context of critiques of contractualist

consideration. Let me label these as follows:[10]

(i) The alienation problem
(ii) The counterproductivity problem
(iii) The exploitation problem

The *alienation problem* is a worry about the link between the values and motivations of actual agents and those of the model agents who represent actual agents in a hypothetical choice situation.[11] The worry is that, if model agents are idealised in certain (perhaps quite radical) ways, actual agents will not be able to identify with them. If actual agents cannot recognise themselves in the model agents, they may not see why the decisions of the model agents are relevant to the decisions the actual agents face in the real world. And this would break the link upon which a prescriptive model relies.[12]

The alienation problem can manifest in different ways. Suppose it is the case that Actual Ian dislikes Italian opera because it is unfamiliar, and he has an aversion to trying out unfamiliar cultural experiences. But suppose that if, for some reason, he did sample Italian opera, he would like it very much. When conceptualising the Model Ian who represents Actual Ian in a hypothetical choice situation, should we suppose that Model Ian likes Italian opera or that he doesn't? Given that the ends of actual agents, and their preferences over possible states of affairs and distributions of rights, will almost certainly be functions of their contingent histories – and that those histories are likely to include instances of irrational decision making – how should we construct ends and preference sets for model agents? Should we take Actual Ian's ends and preferences as fixed and given? What if these include logically contradictory ends or inconsistent preferences?[13] Should we instead eliminate the

accounts of public reason, but some critics do also speak directly to the question of idealisation in contractarian contexts.

[10] I owe the framing of this discussion, and a number of the specific examples, to Gerald Gaus and Zofia Stemplowska.

[11] For a subtle and wide-ranging discussion of this problem, see Thrasher (2019).

[12] Moehler, 2018: 167-168; Vallier, 2019: 99-100.

[13] For a discussion of consistency conditions for agents' preference sets, see §5.2, §5.3, and §5.4.

effects of any rational failings on the part of Actual Ian when constructing Model Ian? If so, Model Ian could end up having very different values. Actual Ian can agree that *if he were* Model Ian, *then he would* value Italian opera. But he isn't, so he doesn't. And now the fact that Model Ian would make decisions on the basis of a value that Actual Ian doesn't share seems to give Actual Ian no more reason to go along with that decision than the fact that Actual Hugo choosing Italian opera would. The alienation problem worries that there is no more relevant connection between Actual Ian and Model Ian than between Actual Ian and Actual Hugo.

The *counterproductivity problem* identifies a different way in which the link between actual agents and model agents can fail. The worry is that, if model agents are idealised in certain ways, they will not suffer from many limitations with which actual agents must reckon (including, for example, weaknesses of will, limited cognitive capacity, limited time for deliberation, etc.). If a prescriptive model says that actual agents (all-things-considered-)ought to do what model agents would do in their place; and if, in circumstances *C*, it would be best to *X* only if one had the capacities of a model agent; then there will be a disconnect between what it is best for a model agent to do in *C* and what it is best for an actual agent to do in *C*. If, in *C*, an actual agent does what a model agent would do in *C*, the actual agent will come a cropper – she will choose an action which is not best for her given her limitations, and which may in fact be disastrous for her.

To illustrate: Suppose Actual Ian is considering whether or not to have wine with dinner. He enjoys wine with dinner, but he plans to go back to work after dinner, and he knows that having too much wine with dinner will lead to him being unable to work effectively after dinner. In order to decide whether or not he has wine with dinner, he consults Model Ian. Model Ian would also enjoy wine with dinner, and also plans to work after dinner (he has the same tastes and ends as Actual Ian). However, Model Ian knows that having more than one glass of wine with dinner will lead to him being unable to work after dinner. Thus, Model Ian would decide to have one glass of wine with dinner – the most wine he can enjoy while still accomplishing his plan of working effectively after dinner. Seeing that Model Ian

would, in the circumstances, have one glass of wine with dinner, Actual Ian decides to have one glass of wine with dinner. Unfortunately, Actual Ian lacks the resolve of Model Ian. After one glass of wine, Actual Ian is feeling overly confident in his capacity to work effectively having had wine, a failing to which Model Ian is not subject. Whenever Actual Ian has one glass of wine with dinner, he always has a second. It is perfectly sensible for Model Ian to decide to have one glass of wine with dinner, because he knows that he will not then have a second glass (his resolve is not inhibited by wine). Actual Ian lacks Model Ian's fortitude. Thus, when Actual Ian does what Model Ian would do in his circumstances, and has a first glass of wine with dinner, Actual Ian puts himself in a position where he will reliably choose to have a second glass of wine, and then will fail to work effectively after dinner. For Actual Ian, doing what Model Ian would do is counterproductive – it fails to achieve his aim of working effectively after dinner. Note that this is true even though we assumed that Actual Ian and Model Ian had identical tastes and ends (and hence we might not be worried about alienation).

I have already touched on an example of the *exploitation problem*.[14] In this case, the question is not whether I (our protagonist, so to speak) should be represented in the contractarian model as having greater (rational, cognitive) capacities or knowledge than I actually have. Rather, the question is, assuming I am so represented to some degree, should the model also represent those actual agents with whom I will actually interact as having greater capacities or knowledge than they will have? This is related to the counterproductivity problem, because what it would be best for Model Me to agree in the hypothetical choice situation depends on whether Model Me assumes the agents with whom I will actually interact have enhanced capacities or merely have their actual capacities. If Model Me assumes the other agents have enhanced capacities when they actually don't, doing what Model Me would do might be counterproductive, as in the second-glass-of-wine-with-dinner example. On the other hand, if Model Me assumes that I will interact with actual agents, with all of their actual limitations, then it is possible that what Model Me would do is use

[14] See the discussion of the idealised poker player in the text, above, and the accompanying n7.

his enhanced capacities to exploit the vulnerable actual agents in various ways. Unless we have some reason to think that Model Me would respond in the same way under either assumption, the output of the contractarian model will vary according to which assumption is made. Under the assumption that other agents have enhanced capacities, we may face the counterproductivity problem. On the assumption that other agents have only their actual capacities, there will be an important inequality in the model, and it is hard to see how anything approximating liberal political institutions can emerge.

The key question, here, is about idealisation: In what ways are model agents supposed to be 'idealised' relative to actual agents? A useful place to begin is with the distinction between moderate and radical idealisation set out by Vallier.[15] In Vallier's account, idealisation is characterised over two dimensions: a rationality dimension and an informational dimension.[16] On the first dimension, idealisation of rationality can be either 'full' or 'bounded'.[17] On the second dimension, idealisation

[15] 2014: 145-180. Vallier's discussion is the most helpfully schematic I have found, but he is picking up a thread from Gaus (2011: 232-258), and indeed in some sense, as Gaus notes (2011: 270), from Rawls (1999: 105-109). Vallier returns to these themes in his later work (2019: 97-103), but his later discussion summarises and is substantially similar to the earlier, and the later text refers us to the earlier, so I will focus on that.

[16] Strictly, Vallier recognises that there may be a third dimension of idealisation, which he identifies as 'reasonableness' (2014: 145, 147-148, 177 n1). Where 'rationality' describes agents' inferential capacities, and 'information' describes their knowledge, 'reasonableness' is supposed to pick out certain dispositions agents are modelled as possessing which are not reduceable to either rationality or information (the disposition to engage in the practice of public justification, for instance). Vallier sets aside any further discussion of reasonableness because he thinks that any public justification account will be trivially committed to idealising for the dispositions covered by reasonableness in at least a minimal way, and he takes no issue with that kind of idealisation (*ibid*.: 148). So long as we understand Vallier's target to be public reason liberals of a Rawlsian inclination, this makes sense. However, the rest of his discussion of idealisation is not limited to Rawlsians. He expressly intends his argument to cover contractarian accounts such as Gauthier's and deliberative accounts such as Habermas's (*ibid*.: 150). If we are talking about idealisation beyond the context of contractualist public reason theories, and specifically, if we are talking about contractarian theories, then a specification of 'reasonable' is not innocuous (*ibid*.: 148). Vallier discusses the reasonableness dimension in more detail in some of his more recent work (2016: 609-611; 2019: 94-97), but he does not consider contractarian accounts in those contexts. For Vallier, what the reasonableness condition does is exclude from consideration at the bar of public reason the demands of grossly immoral or amoral agents. Thus, his account of public reason is less inclusive than mine or, say, Moehler's (2018), though it is far more inclusive than a Rawlsian account of the reasonable. This is in keeping with Vallier's commitment to *moderate* idealisation. For my account of how contractarians should understand the reasonableness condition, see the discussions of the contractarian 'analogue' for unreasonable agents in §2.4.3, §3.1 n19, and §3.2.1 (*supra*), and in §5.5 (*infra*).

[17] 2014: 148. Vallier suggests that, as well as the full vs bounded subdimension of rationality, there are also two others: individual rationality vs collective rationality and interactive rationality vs

of information can be either 'full' or 'adequate'.[18] Given these specifications, we can then say that an account of idealisation is radical if it is committed to full rationality and full information. An account is moderate to the extent that it is committed to bounded rationality and adequate information.[19] As this formulation suggests, while Vallier's sketch sets up radical and moderate idealisations as binary alternatives, he recognises that idealisation can be undertaken over a range from 'full' through varying degrees of 'moderate' to what he calls 'populist' (essentially, the absence of idealisation).[20]

Vallier defends a moderate conception of idealisation against a radical conception (and against populism). He attributes radical idealisation to the bulk of contemporary public justification theories, including Rawls but extending also to contractarians such as Gauthier and deliberative democrats such as Habermas.[21] While I accept some of Vallier's concerns, I will suggest that there are fewer problems with radical idealisation than he thinks.

Vallier argues that radical idealisation is incoherent.[22] The argument goes as follows. First, let us distinguish between *justified* choices and *warranted* choices.[23] A justified choice is "one that a real agent could make given all of the reasoning it has performed *up to the present time* and without violating the constraints of rationality."[24] A warranted choice is "one that would be justified if the agent could

noninteractive rationality. All of the contractualist and contractarian theories with which I am concerned can be specified as either individual-noninteractive-full or individual-noninteractive-bounded, so I omit the complications. See Vallier (*ibid*.: 148-149) for further details.

[18] *Ibid*.: 150. Again, I omit some complications here which are not directly relevant.

[19] *Ibid*.: 151.

[20] *Ibid*.: 157-158, 160.

[21] *Ibid*.: 150.

[22] *Ibid*.: 151-155. Vallier's argument here draws significantly on work by Pollock (2006). In addition to the incoherence argument I discuss in the text, Vallier offers two others (2014: 155-160). However, neither is relevant here. The first is based on integrity considerations. Given that we care about integrity, Vallier argues that moderate idealisation better respects agents' integrity than radical idealisation. But, unlike the contractualists who are Vallier's primary targets, contractarians have no reason to care about integrity. The second argument worries that radical idealisation assumes away (reasonable) pluralism in order to generate unanimous agreement among model agents. Clearly, contractarian accounts don't do this. Instead, contractarians use bargaining theory to generate rational agreement without sacrificing pluralism (reasonable or otherwise).

[23] *Ibid*.: 152; Pollock, 2006: 6.

[24] *Ibid*. Emphasis in original.

complete all possibly relevant reasoning."[25] Radically idealised agents, because they have full rationality and full information, always make warranted choices. In contrast, real agents "cannot and should not always aim to make warranted choices" (at least about anything nontrivial).[26] Real agents at some point have to *act*. Because real agents lack the reasoning resources and information of radically idealised agents they will typically have to act before they have completed *all possibly* relevant reasoning or gathered *all possibly* relevant information. At some point, real agents – even prudent and conscientious agents – will have to determine that they have done *enough* reasoning and gathered *enough* information.[27] While their choices might be justified, they will not be warranted (by definition). Moreover, given how high a bar 'warrant' is, real agents should almost never even aim to meet it.[28]

Thus, Vallier argues, the conclusions of radically idealised agents "will be rooted in reasoning and information that real-world agents are too constrained to utilize. As a result, real-world agents would often be *irrational* to make warranted choices because they would not be able to apply the relevant information in ways that make sense to them."[29] He illustrates this with an example.[30] Suppose Superman rescues John's wife from drowning. As a matter of fact, Superman is Clark Kent, but John doesn't know this (he does not have full information). John, Vallier says, would not be justified in thanking Clark for rescuing his wife, even though thanking Clark would be warranted. Moreover, it would be *irrational* for John to thank Clark, because John lacks any subjectively available reason for doing so. "Therefore, the rational choices of real agents and ideal agents sometimes diverge. While many believe that the rationality of real agents… should approximate ideal rationality, a theory of rational choice for a real agent is only correct if it describes how a real agent should act."[31] Theories of rational choice which employ radical idealisation

[25] *Ibid*.
[26] Vallier, 2014: 152.
[27] On this problem, see the discussion by Elster (1988: 60-62). I owe this reference to David Miller.
[28] *Ibid*. As well as Pollock, Vallier is drawing on a discussion by Gaus (2011: 235-249).
[29] Vallier, 2014: 152. Emphasis in original.
[30] *Ibid*.: 153.
[31] *Ibid*. Compare the discussion of acting on the basis of subjective reasons, rather than objective reasons, in §2.3.1 (*supra*).

will say real agents should act in accordance with their warranted choices, but that is *not* how real agents should act (because it would imply, incoherently, that rational choice requires real agents to act irrationally). So, Vallier concludes, we should reject theories of rational choice which employ radical idealisation.[32]

Vallier's reasoning seems to me to rest on an equivocation and an omission. First, the equivocation. It is true that real agents should not make, or even aim to make, warranted choices, if by 'make warranted choices' we mean 'deliberate and gather information indefinitely, and thus never act'. It doesn't follow, and it is not necessarily true, that real agents should not make, or aim to make, warranted choices, if by 'make warranted choices' we mean 'select the option, from the available options, which you would have selected if you had reasoned without error and had all relevant information'. Real agents should not *follow the procedure* of making a warranted choice (should not deliberate and gather information indefinitely), but they should often *choose in conformity* with warranted choice (should select the option which they would have selected if they had reasoned without error and had all relevant information).

Where does that leave John and Clark? Isn't it still the case that John would be irrational if he thanked Clark, even though thanking Clark would be acting in conformity with (unjustified but) warranted choice? This takes us to Vallier's omission. Consider a different case. Suppose John is travelling in a foreign country where he does not speak the language and is not aware of local customs. One day, he decides to visit a culturally significant site. As he approaches the site, he sees that the woman in front of him, as she reaches the entrance to the site, removes her shoes. Figuring that this woman knows what she's doing, John copies her. Is he justified in doing so? We can suppose that the woman is better informed than John. She has a full grasp of local customs and knows that it is customary to remove one's shoes before entering this particular site. John is warranted in removing his shoes

[32] Vallier supplements this argument with an argument which relies on the distinction between choice over actions and choice over plans (2014: 153-154). His remarks are sensible, but they do not add substantially to his critique of radical idealisation.

(assuming he wishes to conform to local customs, etc.). But I argue that he is also justified in removing his shoes. John correctly intuits that the woman knows more than he does. He need not have any idea *why* she is removing her shoes – her reasoning might be thoroughly opaque to him. But he can still use her as a model for his own action. He is justified in removing his shoes *by the fact that* he is copying her, though he clearly doesn't act for the same (subjective) reason as she does. John, in this case, does nothing irrational. What is missing from Vallier's Superman example is a model for John's actions which can (subjectively) justify him acting in accordance with warranted choice.

More generally, what is missing from Vallier's story is that we can have reason to act in conformity with the conclusions of a model agent even if we do not have the characteristics (full rationality, full information) of the model agent. Now, of course, John has to have good reason to think that the woman in front of him *is a good model* for him. Perhaps she isn't. Perhaps she is removing her shoes, not because she is following some custom, but because they have become uncomfortable. If so, he is going to feel quite foolish when he gets inside to find everyone else in shoes. Similarly, in order to be justified in acting in conformity with the conclusions of a contractarian model, real agents need to have reasons to think that the model is a good one. But my point here is that they can have such reasons without being model agents. There is nothing irrational in real agents conforming to the choices of model agents, even if those model agents are radically idealised, if the real agents are conforming *because* they are using the radically idealised agents *as* models. The point of a model is to bridge the justificatory gap between the choices of model agents and the reasons of actual agents.

We should take care not to move too quickly from the case of John the Tourist to an acceptance of radical idealisation in a contractarian model. For one thing, John copies the actions of a real agent whom he takes to be a model. It is much less obvious how one can 'copy' the actions of hypothetical agents – these must be constructed since they cannot actually be observed. For another, as I have already noted, real agents need reasons to think that some particular model is a good model.

When John enters the site he is visiting, there will be an external check on whether the woman he is copying has been a good model for him. This is true of empirical models generally – one can be confident one has hit upon a decent model of aerodynamics when one's aeroplanes no longer fall out of the sky. It is far from clear how a similar check can be available if one's model is concerned to identify principles of justice to regulate the basic structure of society, say.[33] What check can I have that a model with hypothetical agents, employing radical idealisation, is a good model for me, except for my direct assessment of the conclusions of the model (which, if I am not fully rational or fully informed, I will not reach myself)?

These are sensible concerns. While I don't think they give us reasons to reject radical idealisation as generally as Vallier seems to want to argue, they should prompt us to be more cautious in entertaining radical idealisation than many mainstream theories have been. One potentially promising way of proceeding is to follow Vallier in disaggregating different aspects of idealisation. Vallier has already distinguished the rationality and informational dimensions. Within the rationality dimension, I think we might also distinguish three components. First, there are certain formal consistency requirements. Second, there is the capacity to make error-free inference. Third, there is what we can call 'processing power' – the capacity to apply one's reason as quickly as necessary to any amount of information (or, which is the same thing, to have as much time as needed to apply one's reason to any amount of information).[34]

It is useful to separate out these four elements of idealisation (three rationality-related, one informational) because they don't all look equally suspect given the concerns identified above. That is to say, it should be relatively easy for a real agent to see why a model is better if it includes the consistency requirements and error-

[33] Moreover, attempts by contractualists to impose such checks in the form of 'stability' raise a host of problems. See Thrasher's provocative discussion (2019).

[34] Compare Vallier (2014: 149): "The fully rational agent reasons according to the standard axioms of decision theory (completeness, asymmetry, transitivity, etc.) and make[s] perfect inferences based on her beliefs and values. Bounded rationality relaxes these requirements. Accounts of bounded rationality model reasoning based on the constraints faced by real-world agents, such as limited processing power and high information collection costs."

free inference than in their absence. While these things certainly aren't uncontroversial, and precisely how one specifies them may be important, what they have in common is that they can be contemplated at a very ordinary scale. The capacity to see what's wrong with reasoning that violates consistency or valid inference is, comparatively speaking, quite an ordinary capacity in any particular instance. What makes idealisation on these elements radical is the iteration of them over multiple, interlocking instances. Real agents make mistakes at such scales, where ideal agents won't. But real agents can grasp the point of avoiding inconsistency and invalidity at the scale of everyday instances of reasoning and can then extend this idea generally.

The same may not be true for the processing power capacity, and certainly isn't true for information. Real agents don't have the same imaginative acquaintance with unlimited processing power or omniscience that they can have with consistency and error-free inference simply by generalising from particular instances of consistency and error-free inference. Not all elements of radical idealisation are equal, then – some are more suspect than others.

Vallier defines radical idealisation as being committed to full rationality and full information. He instead favours moderate idealisation, and thus bounded rationality and 'adequate' information. I have suggested that Vallier's worries about radical idealisation are less persuasive than they first appear. However, I acknowledge that they are not without foundation. For the purpose developing my own model, I will propose 'quite radical' idealisation. I intend this to idealise more in some ways than Vallier would like, but also to fall short of the fully radical idealisation he criticises most persuasively.

My model of rational agents will be specified as follows. In order to employ the formal foundations I wish to use, I need to insist upon the consistency requirements. However, as I've just argued, idealising in this way – even radically – does not raise significant costs in terms of the acceptability of the model to real agents. For a similar reason, I think we can assume error-free inference. Processing power and

information require more moderation.[35] In §5.3, I allow agents sufficient processing power to achieve a complete preference ordering, for the same reason that I have given in defending idealised consistency and inference – as an iteration from an easily accessible specific case. However, agents' capacity for information processing should not be taken to include infinite time to *revise* their preferences in light of further reflection (for the reasons Vallier and Pollock cite). Further, as I will set out my model, the agents need to know *enough*, but they do not need to know everything. As I discuss in §7.4, model agents can respond rationally to the fact that they anticipate not having full information. Note also that my model includes no veil of ignorance, or any similar device, so agents are not 'idealised' in the sense of having *less* information either.[36]

For the specific purposes of a contractarian model, one particularly thorny problem is how to conceptualise *patience*.[37] In one sense, patience seems to be an aspect of rationality. Certainly, we might think highly impatient agents are to that extent irrational. The capacity to delay gratification is one of the key cognitive abilities we associate with maturing from a (less rational) child to a (more rational) adult. On the other hand, it is almost axiomatic in economic theory that rational agents discount future gains in some way (for a variety of reasons, some uncontroversially rational – those associated with uncertainty, for instance). Economic theory typically takes an agent's discount rate to be simply given by his preferences. But patient agents can have advantages in bargaining situations, and so *prima facie* we have good reason to conceive of model agents as more patient than actual agents. Patience lies at the junction of preferences and rationality, and no matter how a theory resolves its conception of patience it will leave itself exposed to objections from one side or the other.[38] Given that the stakes are high in the contractarian model – what is being determined is the nature of political society – and that the model is prescriptive

[35] *Ibid*.: 161-162.
[36] See *ibid*.: 150-151.
[37] Thoma, 2015: 3348. See §6.4 n35, the introductory remarks in Chapter 8, and §8.4 n53, *infra*.
[38] It might be possible to give a more fine-grained analysis of impatience, in which one attempts to distinguish between those aspects of impatience which result from things like tiredness or frustration or similar frailties, and aspects of impatience which reflect a more purely evaluative attitude to gratification. One might then argue that it is appropriate to idealise away the former, while retaining the latter. I do not have the space to develop such an account more fully here.

rather than descriptive, in the discussion to follow I will err on the side of idealisation in conceptualising patience.

With this discussion of idealisation in hand, let me return to the three problems I identified earlier. Consider first the problem of alienation. Recall that this was the worry that the values and motivations of model agents and those of actual agents might come apart, such that actual agents might not identify with the choices made by model agents and consequently might not be motivated to actually act on the basis of the fact that the model agents have chosen in a particular way. The first thing to note is that the idealisations my model makes are broadly procedural rather than substantive. Where the model does idealise, in order to eliminate inconsistencies or errors in inference on the part of actual agents, say, my claim is that these changes would be endorsed by the actual agent in question upon a reasonable amount of thoughtful reflection.[39] I assume that model agents have access to *at least* the information that actual agents have. There is no veil of ignorance, or any similar device, placed between the values of actual agents and their model representatives. However, I also do not assume that model agents have access to counterfactual information of the kind represented by the Italian opera case. I allow that model agents may have *some* more information than actual agents, because I allow that model agents may have filled out inferentially some of the implications of actual agents' beliefs and values. But, again, this is constrained to what is possible upon a reasonable amount of thoughtful reflection, given moderate processing power. Finally, while I assume, for the sake of argument, that many agents have a disposition to engage in agreed patterns of coordination as an alternative to violent conflict, because without this assumption my discussion would be wholly academic, I do not make any stronger assumption of reasonableness than this (I do not assume anything as strong as a public justification principle, understood in the standard way).[40]

[39] See Gaus's discussion of a 'respectable amount' of reasoning (2011: 244-257). As Gaus notes (*ibid*.: 254), it is hard to give very determinate boundaries to this idea, at least in the abstract. It again seems plausible to me to say that consistency and error-free inference do not require *excessively* demanding reflection, so long as we continue to consider particular instances of reasoning rather than more extended forms of reasoning.
[40] Vallier, 2014: 24.

While I have described my conception of idealisation as 'quite radical' on some dimensions, it is from the point of view of the alienation problem that it skews most moderate.[41] Contractarian theories have plenty of weaknesses, but one of their principal strengths is that they take the existing ends of actual agents more or less as given. Contractarians do not require that actual agents imagine their ends to be wholly other than they are for the purpose of modelling their hypothetical choices, but simply that their reasoning is procedurally enhanced in certain ways. This kind of idealisation is not costless, but the payoff is worth the price.

If we think of the alienation problem as raising two kinds of question for a contractarian model – a question about values and a question about motivation – the questions can be answered as follows. Actual agents should almost always be able to recognise their values in model agents. The only time model agents and the actual agents they represent would have different values is when the process of eliminating inconsistencies and errors in inference has led to a modification of the preference profile of the model agent. Even then, the values should not be wholly alien, since eliminating an inconsistency between two of an actual agent's ends should leave one of those ends intact. Given this, there also shouldn't be a problem of motivation. The choices of the model agents will always be underpinned by some value to which the actual agents are committed, and which is, for them, motivating.

There are some bullets to be bitten in following this approach. In taking the values of actual agents as given, I am accepting the contingent origins of those values. In particular, there's nothing to say that the values actual agents have, and which model agents will inherit, are not the result of injustice or abuse or false consciousness. If one's aim were to use a social contract model to identify morally sound principles of justice, this would be a problem, and greater idealisation would be appropriate.[42]

[41] The possible exception here follows from my remarks about patience. We might worry that the alienation problem will be acute in cases where ideally patient model agents are representing highly impatient actual agents. Still, the response in the discussion of counterproductivity, below, goes some way to addressing this worry. Consider an extreme case – a drug addict at risk of extremely painful withdrawal. The physical fact of withdrawal seems to be something that can be incorporated into the deliberations of an ideally rational model agent in the same way a possible failure of resolve could be.

[42] Compare Arneson's instructive discussion of malformed preferences (1990a: 167-170).

But the contractarian purpose has a more practical orientation. With that purpose in mind, we must take agents largely (if not entirely) as they are.

What about the counterproductivity problem? Recall that this problem arose because actual agents and model agents have different capacities, and sometimes what it is best for one to choose depends on what capacities one has. The best choice for a model agent, with model capacities, might not be the best choice for an actual agent with actual capacities, and so there is no straightforward line that can be drawn from the choices of model agents to the reasons of actual agents.

In part, the counterproductivity problem is solved simply by the clarifications already made. Consider the second-glass-of-wine-at-dinner case. As I have described my model agents, it is no part of the description of their idealisation that they do not suffer from failures of resolve. Model Ian, then, need not assume that he differs from Actual Ian in this respect. The counterproductivity problem doesn't arise, because Model Ian is just as worried about failures of resolve as Actual Ian.[43]

This response limits the scope of the counterproductivity problem but perhaps it does not resolve it entirely. Suppose our case is one in which the relevant capacity *is* a capacity in respect of which Model Ian and Actual Ian differ, and is a respect in which my model requires Model Ian to be enhanced relative to Actual Ian (so, perhaps the error-free inference requirement, or even just the adequate information requirement, which is still an idealisation, albeit a moderate one). Now it seems the counterproductivity problem might retain its force. Nonetheless, cases such as this look quite odd. Suppose Actual Ian is in a situation in which he must make some kind of inference, and that his inferential capacities are imperfect, such that he sometimes makes inferential mistakes. If, in such circumstances, he consults Model Ian, who makes no inferential mistakes, and then does what Model Ian would do, how can the result be counterproductive? Model Ian's reasoning is, by definition,

[43] As far as I can see, this response is extensionally equivalent with certain ways of specifying 'ideal advisor' models of idealisation (Smith, 1994: 151ff). To the extent there are differences, it seems to me that fewer problems of representation arise when proceeding as I do in the text. Compare Vallier, 2014: 154-155.

maximally productive (at least insofar as it maximises *expected* utility). Of course, it remains true that Actual Ian may not be able to determine what Model Ian would do, or he may know but fail act accordingly. But those are not problems of representation, they are inabilities or failures to apply the model.

This still might be thought too quick. Consider: Sometimes, one of the things we have to decide is whether or not to put ourselves in the position of having to make further decisions. Model Ian has no fear of making future mistakes, because his inference is error-free. But Actual Ian can have no such confidence – perhaps he shouldn't follow a path which will expose him to increasing probability of error. Extrapolated to the level of societies and states, this might lead us to Oakeshottian or Hayekian concerns about the efficacy or wisdom of central planning or rationalist social reform.[44]

There are two things to say about this. First, constructed in this kind of way, the case looks like a return to the second-glass-of-wine-at-dinner problem. Just as we need not construct Model Ian such that he has a resolve Actual Ian lacks, there is no requirement that we idealise Model Ian such that his error-free inference is a *persistent* feature – he can be ideally rational *now* (in the relevant respects), while anticipating having degraded capacities at some later time.[45] And then the response is the same as before – the counterproductivity problem can't arise, either because model agents anticipate the limited capacities of actual agents, or because the idealisation of model agents is maximally productive by definition. The second thing to note, then, is the general response to conservatism of an anti-rationalist bent: The presumption in favour of the *status quo* is not indefeasible. The possibility of error cuts both ways. Sometimes inaction is more costly than (potentially erroneous) action.

[44] Oakeshott, 1991: 5-42; Hayek, 1945. Compare Mill, 1974: 151.
[45] This strikes me as a perfectly mundane feature of human psychology, or at least of anyone who enjoys wine. Again, put like this, I see no relevant difference between this response and an ideal advisor account (see n43), but strictly this is still a case of doing what the (more precisely, realistically specified) model agent would do, rather than what the model agent would advise the real agent to do.

Finally, consider the exploitation problem. It is worth unpacking this problem a little more, since it introduces an additional layer of complexity not present in the more basic (parametric) examples of the counterproductivity problem. There are three possibilities to consider:

(i) Model Me is bargaining with Model Others and generating a prescription for a world with model agents.
(ii) Model Me is bargaining with Actual Others and generating a prescription for a world with actual agents.
(iii) Model Me is bargaining with Model Others and generating a prescription for a world with actual agents.[46]

Option (i) raises the counterproductivity problem, and so we should set it aside. However we decide to construe the agents engaged in bargaining, the model should issue a prescription which accommodates the limitations of actual agents in the real world. However, this opens up the possibility of exploitation.

The exploitation problem worries that, if one wishes to avoid exploitation, one must assume model agents; but that if one wishes to avoid counterproductivity, one must *not* assume model agents. The point of distinguishing Option (ii) and Option (iii) is to defuse this claim. Both options avoid counterproductivity, because they assume that the prescriptions of the model will be constructed with real, limited agents in mind. But Option (iii) avoids counterproductivity while also circumscribing the scope of possible exploitation. It idealises other agents for the purposes of bargaining – hence, they are not vulnerable to exploitation (at the point of bargaining). But it does not idealise other agents for the purposes of prescription. The model agents engaging in hypothetical bargaining will bargain in the knowledge that the real agents they represent will lack their enhanced rational capacities.

[46] The fourth logical possibility is Model Me bargaining with Actual Others and generating a prescription for a world with model agents. But this isn't a plausible alternative. More generally, see Moehler's illuminating discussion of his distinction between 'logics of thinking' and 'logics of general real-world behaviour' in the context of his empathetic contractor position (2018: 110-112).

I take it this response disarms the immediate dilemma. Still, Option (iii) is at this point simply a logical possibility. Why *should* we construct the model along these lines? From the point of view of any particular agent, what is it that is gained by representing other agents as more capable than they are, and thereby foregoing the potential gains of exploitation?

At this point, it is worth returning to the argument of Chapter 2. There, I argued that real agents (all-things-considered-)ought to give one another prudential reasons when interacting politically, because such reasons will be motivating for the people to whom they are addressed. The point of building a contractarian model in Part II is to work out what set of political institutions an agent is justified in endorsing, given that she acts on the basis of her prudential reasons (the reasons she will be offered by other agents, following my reasoning in Chapter 2). So, the story from Chapter 2 is that I tell you that you (all-things-considered-)ought to coordinate with me on a set of institutions *S* because that is what Model You would agree when bargaining with Model Me. No purpose would be served by me saying, instead, that you (all-things-considered-)ought to coordinate with me on a set of institutions *T*, because that is what Actual You would agree when bargaining with Model Me. What Actual You would agree is not normative for you. Now, you might reply that you are not interested in what Model You would agree when bargaining with Model Me – you want to know what Model You would agree when bargaining with Actual Me, because you see no reason to forego the possibility of Model You exploiting Actual Me. But, again, what purpose is served by this? *That* way of modelling our interaction would not be normative for *me*, and you need to motivate me as much as I need to motivate you. What about third parties? Perhaps you and I can agree that we should be idealised in the model, but can also agree that we should model others as their incapable selves. But the problem just repeats itself. To the extent we need the willing coordination of third parties (and I assume we do), we must also model them as idealised (at the point of bargaining). Otherwise, we won't have given them a reason to coordinate.

All of that said, I do not mean to exclude the possibility of exploitation entirely. The

reasoning I give in Chapter 2 relies on agents *believing* themselves to have been given reasons to coordinate in a particular way.[47] If Actual Me can fool Actual You into accepting a model of prudential reasons that is rigged in my favour, so much the better for Actual Me. Equally, if I can convince you that I am your god incarnate and you owe me unquestioning obedience, that is what I (prudentially-)ought to do. However, in §7.3 I give some reasons I think we have to *theorise as if* this were not the case. First, because it is sensible not to rely on one's capacity to indefinitely hoodwink others.[48] Second, as what I call a 'first-cut' idealisation – a simplification of the theory pending further refinement. Third, because I argue that the cognitive capacities necessary for understanding the model I present are in fact relatively moderate, and will probably be within the grasp of many fellow citizens. After all, it will not take much close inspection on your part to notice that I have not modelled you in the same way I have modelled myself. Option (iii), suitably specified, then seems to me to be a defensible characterisation of the contractarian model, avoiding both (theoretically problematic) exploitation problems and counterproductivity problems.

Let me offer a final thought. Throughout this discussion of idealisation, I have mostly been concerned to defend idealisations which (in sum, if not one by one) fall between the fully radical and the (to my mind) excessively moderate. Where I have leaned more to the radical, I have mostly defended my conception against objections, rather than giving a positive case for it. Before concluding, it may be helpful to recall what it is that is *attractive* about radical idealisation.[49] Consider the most extreme alternative to radical idealisation – what Vallier describes as 'populism' – the absence of any idealisation whatsoever. Eliminating idealisation altogether from a model renders the model entirely descriptive or predictive.

[47] §2.3.1.
[48] Compare Hobbes, in *Leviathan*, Ch. 15 (1996: 102): "[W]hen a man doth a thing, which notwithstanding any thing can be foreseen, and reckoned on, tendeth to his own destruction, howsoever some accident which he could not expect, arriving, may turne it to his benefit; yet such events do not make it reasonably or wisely done… He therefore that breaketh his Covenant… cannot be received into any Society… but by the errour of them that receive him… which errours a man cannot reasonably reckon upon as the means of his security…"
[49] Compare Vallier, 2014: 145-146, 149-151.

Prescriptive models *must* idealise. *Justification* inheres in idealisation.[50] Each step we take away from radical idealisation comes with a justificatory cost. For reasons tied to worries about alienation, contractarians will buy much more moderation than contractualists will. But still, we should only pay the costs of moderation if we're getting our money's worth.

With these various remarks on idealisation in mind, I proceed as follows. In order to address the problem of insufficient justness discussed in Chapter 3, I intend to show that my contractarian model prescribes liberal political institutions. Thus, I will need an account of liberalism. I set this out in Chapter 4.

In Chapters 5 and 6, I develop the contractarian model. In Chapter 7, I argue that, given the characterisations I have taken of rational agents, their circumstances, and rational bargaining – and given that we assume agents have equal 'bargaining power' – the agents in the contractarian model would agree to coordinate in a manner which meets the standard for liberal political institutions specified in Chapter 4.

It will remain to show that the assumption of equal bargaining power can be justified. That will be the task of Chapter 8, where I will also return to the problem of practical stability. In combination, then, the remaining chapters will establish a contractarian argument for liberal political institutions.

[50] Vallier, 2019: 99, 100.

Chapter 4

In this chapter, I defend a conception of liberalism. My discussion is structured around two different ways of understanding liberalism: the authority-based conception and the scope-based conception. The authority-based conception is a constructive approach to liberalism.[1] It defines as liberal whatever political institutions would be chosen by suitably idealised agents – agents who are rational, free, and equal – under suitably idealised conditions. The scope-based conception is a more conventional approach. In defining liberal political institutions, it draws on our conventional understanding of institutions which are secular, under which citizens have equal status, and under which citizens enjoy a wide range of freedoms in their everyday lives.

In §4.1, I argue for the authority-based conception. One of the reasons I favour the authority-based conception is that it gives us a mechanism for systematically evaluating and revising our conventional understanding of liberalism. Nonetheless, I concede that there is a limit to the extent to which an authority-based conception can depart from our pre-theoretic understanding of liberalism without becoming something conceptually distinct from liberalism. Thus, I conclude that a set of political institutions qualifies as liberal if and only if (i) it would be chosen by suitably idealised agents, *and* (ii) also falls within a reasonable range of our conventional understanding of liberalism.

In §4.2, I argue that a contractarian theory can qualify as an authority-based conception of liberalism. This being so, if the institutions prescribed by the model fall within a reasonable range of our conventional understanding of liberalism, we

[1] For a recent critical discussion of constructivism in political philosophy, see Thrasher (2019). As a general rule, constructivist approaches get into trouble when they make exaggerated claims for the parsimony of their premises. Rawls is sometimes guilty of this (2005: 94, 116, 126-129). So long as one does not attempt to go without *any* foundations, most of the trouble can be avoided.

will have a contractarian argument for liberal political institutions.

§4.1 Two conceptions of liberalism

Suppose a set of political institutions is properly characterised as liberal if and only if citizens would have sufficient reason to choose it under conditions in which they are free and equal.[2] Note two aspects of this conception of liberalism. First, on this account, the central concern of liberalism is the justification of political institutions. The conception is expressed in terms of a hypothetical choice, but the link between hypothetical choice and justification should be clear.[3] Why not say, then, that political institutions are liberal if and only if they are justified? As I noted earlier, there are two senses of justification – justification *simpliciter* and justification *to* some person.[4] Liberalism is not concerned with justifying political institutions to just anyone. It is concerned with justifying political institutions to, and hence by the lights of, free and equal agents. The hypothetical choice construction makes explicit this second aspect of the conception.

Why should we think that liberalism has anything to do with justifying political institutions to free and equal agents? It is uncontroversial that liberalism is connected in *some* way to the concepts of freedom and equality.[5] What needs defence is the claim that justification to free and equal agents best captures the link between liberalism and freedom and equality.

---

[2] The clearest statement of this conception is given by Jeremy Waldron (1987). See also Jaggar, 1983: 33; Hampton, 1997: 180; Horton, 2010: 438; McCabe, 2010: 3-7; Rossi, 2010: 21-22; Ryan, 2012: 37-38; Gaus, Courtland, and Schmidtz, 2018: §1.1.

[3] §1.2.

[4] §2.4.1 n53. Compare Scruton, 2001: 182-194.

[5] Jaggar, 1983: 33; Raz, 1986: 2; Macedo, 1991: 2, 9-10; Vincent, 1992: 32, 37, 41-42; Hampton, 1997: 170-173, 179-182; Rawls, 1999b: 14; Gaus, 2000: 159-160, 166-168, 200; Gray, 2000: 47; Rawls, 2001: 5-8; Gaus, 2003: 1, 3; Rawls, 2005: 6-7; Held, 2006: 59, 61-62; McCabe, 2010: 4, 126; Rossi, 2010: 23; Quong, 2011: 2-3, 14-15; Stilz, 2009: 7-9; Ryan, 2012: 23, 28, 30, 40; Zellentin, 2012: 20-21; Gaus, Courtland, and Schmidtz, 2018. Compare Dworkin, 1985: 181-204; Freeden, 1996: 159-160, 231, 241-247, 430-433.

This conception of liberalism can be traced to the Enlightenment.[6] Rather than seeing subjection to political authority as natural or divinely ordained, the starting point of this tradition is the idea that individuals (i) are 'naturally' or inherently free, and (ii) possess a capacity for reason.[7] 'Free', in this context, means 'not properly subject to authority'. That all people alike are both free and rational means that people are, in a basic sense, equal.[8] In the social contract theories of the period, this equal status was expressed using the device of the state of nature, a state prior (logically, if not historically) to the formation of political institutions. The basic problem of Enlightenment political thought is how to reconcile the inherent freedom and rationality of human beings with the existence of political authority. Given that people are not subject to authority by default, political institutions stand in need of some special grounding.[9] That grounding is provided by the idea of rational consent. If people rationally choose to submit to political authority, authority and freedom are reconciled.[10]

Thus, we have an account which connects freedom, equality, and rational choice (choice in accordance with reason). Formulating the conception in terms of hypothetical consent, and hence justification, rather than in terms of actual consent, is potentially problematic.[11] But the role that rationality plays in the Enlightenment theories opens up the space for such a move.[12] The normative significance of consent

[6] Waldron, 1987: 134-136; Hampton, 1997: 180-181; Gaus, 2000: 56-57; Gaus, 2003: 1-5; Held, 2006: 58-65. Compare Ryan, 2012: 8-9, 14. My discussion here takes liberties which historians of political thought will find problematic. In mitigation, I would note that my purpose is to articulate the way in which a particular strain of contemporary liberalism understands its historical antecedents, rather than to engage in serious intellectual history. I am grateful to Sam Kiss for helpful comments on this section.

[7] See Hobbes, in *Leviathan*, Chapters 3-5, 13-14, and 17-18 (1996: 20-37, 87, 91-94, 120-121); Locke, in *The Second Treatise of Government*, Chapters 1-2, 4, 6, and 8 (2005: 17-23, 26-27, 39-50, 58-71); Hume, in "Of the Original Contract" (1948: 357-359); Rousseau, in *The Social Contract*, Book I, Chapters 1-5 (1968: 49-59). See also Jaggar, 1983: 28, 33; Hampton, 1997: 181-182; Held, 2006: 59; Stilz, 2009: 7-8; Ryan, 2012: 28-30.

[8] Jaggar, 1983: 29; Hampton, 1997: 181-182; Gaus, 2000: 159; Rawls, 2001: 196.

[9] Eccleshall, 1994: 32-33; Scruton, 2001: 186-187; Enoch, 2015: 114-115.

[10] Gaus, 2000: 159; Scruton, 2001: 186-187.

[11] Waldron, 1987: 138-146.

[12] Hampton, 1997: 180-181. Compare Hobbes, *Leviathan*, Chapter 14 (1996: 92-94, 96-99); Locke, *The Second Treatise of Government*, Chapters 6 and 9 (2005: 40-43, 73); Hume, "Of the Original Contract" (1948: 359); Rousseau, *The Social Contract*, Book I, Chapter 8, and Book II, Chapter 3 (1968: 64-65, 72).

in such accounts is that it is an expression of one's capacity for reason. If the only choices which are normatively significant are those which track reason, then hypothetical rational choice is a valid substitute for actual choice.

Because the central concern of this sort of account is the justification of political authority, I refer to this conception of liberalism as the *authority-based* conception. I take it the authority-based conception is, on its face, a plausible conception of liberalism, broadly understood. It is grounded in a recognisably liberal tradition and it accomplishes the basic task of bringing freedom, equality, and political institutions into some kind of coherent theoretical relationship with one another.[13] Given the role that hypothetical choice plays in contractarian theory, the authority-based conception is obviously an attractive conception of liberalism for anyone hoping to construct a liberal contractarian theory. The fit between contractarianism and the authority-based conception of liberalism will be quite natural.

I will argue for the authority-based conception of liberalism in two ways. First, I will present a set of criteria which it is reasonable to think a conception of liberalism should satisfy. The authority-based conception satisfies these criteria. Second, I will compare the authority-based conception with what I take to be the most plausible alternative – what I will call a scope-based conception – and argue that the authority-based conception has notable advantages.[14] My aim is not to produce anything so ambitious as a definitive argument in favour of the authority-based conception. I will simply claim that an authority-based understanding of liberalism provides a defensible platform from which to develop the rest of my argument.

Liberalism is a theory of politics which organises a variety of concepts into a coherent structure.[15] Different conceptions of liberalism organise its component concepts in different ways (and employ different conceptions of those concepts).

[13] The emphasis here must be on liberalism *broadly understood*. On the status of Hobbes and Rousseau in the liberal tradition, for instance, see Gaus, Courtland, and Schmidtz (2018: §1.1).
[14] For a similar distinction between the justificatory and substantive elements of liberalism, see McCabe (2010: 3-7, 135).
[15] Gaus, 2000: 26-43.

There is inevitably a degree of arbitrariness about drawing the boundary between liberal and non-liberal theories, but the following six concepts plausibly represent necessary (and jointly sufficient) components of any theory we would want to describe as liberal. First, obviously, freedom will play a central role. Second, equality will be similarly important. I have already touched on these elements in introducing the authority-based conception.[16]

Third, then, is the idea of 'neutrality'.[17] Neutrality is a concept that has been particularly prominent in recent philosophical discussion of liberalism. The term itself comes with some unfortunate baggage.[18] I use it here to refer to the idea that the state is, or (in-some-sense-)ought to be, constrained from taking sides in certain disputes between citizens (usually, disputes about what constitutes a valuable way of life). As such, I take conceptions of neutrality to be contemporary glosses on, or extensions of, the older liberal idea of toleration.[19]

When I have mentioned liberalism earlier, I have used freedom, equality, and neutrality as a shorthand for describing its commitments. Looking more closely, three more of its features deserve attention. The fourth component of liberalism is the idea of individuality or individualism.[20] Several different ideas fall under this heading, but I will highlight two kinds in particular. On the one hand, we have a substantively normative idea – roughly, the thought that non-conformity and diversity are positively valuable characteristics of citizens in a society.[21] On the other

[16] For supporting references, see n5.
[17] Nozick, 1974: 33; Ackerman, 1980: 10-12; Dworkin, 1985: 191-192; Gauthier, 1986: 341; Larmore, 1987: x, 42-68; Narveson, 1988: 8; Arneson, 1990a: 194; Arneson, 1990b: 446-451; Larmore, 1990: 341; Larmore, 1996: 121-134; Hampton, 1997: 173-180; Rawls, 1999a: 80-81; Gaus, 2003: 1, 15-16; Heywood, 2003: 36-39; Rawls, 2005: 190-195; Held, 2006: 59; McCabe, 2010: 4; Quong, 2011: 1-2; Zellentin, 2012: 1-11; Vallier, 2013: 646. For surveys of the relevant literature, see Raz (1986: 107-162); Sher (1997: 1-3, 20-44); Quong (*op. cit.*: 15-22).
[18] Rawls, 2005: 191.
[19] Rawls, 1985: 228-231, 249; Larmore, 1987: 51; Gray, 2000: 2-3; Gaus, 2003: 15-16; Rawls, 2005: xxi-xxviii; Held, 2006: 59; Scheffler, 2010: 279-280; Zellentin, 2012: 42-44. For conceptions of liberalism instead committed to toleration, see Raz (1986: 403-404, 406-407), Macedo (1991: 278-279), Hurka (1993: 148-160), Sher (1997: ix), Ryan (2012: 7-8, 22, 30-32, 35).
[20] Jaggar, 1983: 28-29, 33; Ripstein, 1987: 134; Vincent, 1992: 29, 32; Freeden, 1996: 144-148; Hampton, 1997: 172, 180; Rawls, 1999a: 3, 24-26; Gaus, 2000: 61-62, 69, 170, 200-201, 204; Gaus, 2003: 1; Heywood, 2003: 28-30; McCabe, 2010: 150 n49; Ryan, 2012: 8-9, 23, 35-37; Gaus, Courtland, and Schmidtz, 2018: §3.4.
[21] Most famously, see Mill's discussion in *On Liberty*, Chapter 3 (1974: 119-140). Compare Muldoon

hand, we have something more like a methodological or metaphysical idea.[22] The thought here is that the relevant units of analysis for political theories are individual people, rather than groups or nations or communities. The fifth component of liberalism is the concept of rationality or reason.[23] Finally, sixth, is universalism – the idea that political theories are (at least in some sense) unbounded by temporal or cultural contexts, and apply in all times and places, and to all groups of people.[24] Not everyone would agree that each of these six elements is a necessary component of liberalism. Some would wish to add different elements to the list.[25] But, as a characterisation of liberalism, this set of commitments seems reasonable enough.

It should be evident that the authority-based conception qualifies as a conception of liberalism by this measure. It understands freedom and rationality as characteristics of citizens which ground a requirement that political authority be rationally justified. In requiring a justification to be given to each citizen, equally and individually, it incorporates liberal commitments to equality and (methodological) individualism. Importantly, in the requirement that political institutions be justified *to* citizens, rather than justified *simpliciter*, there is a commitment to a version of neutrality: If political institutions must be justified to each citizen, and if the citizens disagree about values, the justification of the institutions cannot be based solely on some sectarian account of value. Finally, insofar as we take freedom, equality, and rationality to be characteristics of all people, in all times and places, the authority-based conception claims universal application.

(2016).

[22] See Heath (2015) for a discussion of methodological individualism and its relation to 'atomistic', metaphysical individualism.

[23] Jaggar, 1983: 28-30, 33; Waldron, 1987: 132-133; Freeden, 1996: 148-150; Hampton, 1997: 180-182; Rawls, 1999a: 123-126; Gaus, 2000: 54, 56-60; Rawls, 2001: 196; Gaus, 2003: 1-3; Heywood, 2003: 33-34; Rawls, 2005: 48-54; McCabe, 2010: 5.

[24] Jaggar, 1983: 30-32, 35; Freeden, 1996: 241; Hampton, 1997: 181; Gaus, 2003: 1-3; Stilz, 2009: 9.

[25] Additional candidates include progress (Freeden, 1996: 145, 147; compare Ryan, 2012: 24-25), sociability (Freeden, 1996: 150), the common good (*ibid*.: 151), constitutional or limited government (Jaggar, 1983: 34; Vincent, 1992: 48; Freeden, 1996: 151-152; Held, 2006: 59; Ryan, 2012: 24, 29, 39-40), democracy (Vincent, 1992: 44-45; Freeden, 1996: 154-159; Hampton, 1997: 179-180; compare Larmore, 1990: 359 n15; Ryan, 2012: 24, 39), and private property (Jaggar, 1983: 34-35, Vincent, 1992: 43, 45-47; Freeden, 1996: 163-165; Held, 2006: 59; Ryan, 2012: 24-25, 33-34). Of these, constitutionalism and democracy strike me as most promising, but I believe they can be accommodated as further specifications of freedom and equality.

So, it's reasonable to describe the authority-based conception as a conception of liberalism, although it clearly represents just one way to organise the various components. How does it compare to alternative conceptions? I cannot conduct an exhaustive comparison of candidate conceptions here. But in order to defend my use of the authority-based conception, I will compare it with what I take to be the most plausible alternative: a scope-based conception.

One reaction to the authority-based conception might be something like this: The authority-based conception misses the point. Liberalism isn't defined by abstruse hypothetical reasoning. It doesn't have any particular link to the justification of authority. Rather, political institutions are liberal when they don't interfere with citizens' self-regarding actions; when they treat each citizen as equal before the law; when they are officially and effectively secular. More generally, a set of political institutions is properly characterised as liberal if and only if, under the institutions, the scope of each citizen's freedom is sufficiently wide, each citizen has freedom of an equal scope, and the institutions are constrained by a principle of neutrality or toleration. Liberalism, on this view, is concerned with substantive questions about how the state interacts with citizens; it is not constitutively 'procedural' in the way the authority-based conception imagines.[26] Call this a *scope-based* conception of liberalism.[27]

Although it will arrange the various concepts in a different way, a scope-based conception can incorporate the six components of liberalism I identified above. It will contain substantive requirements relating to freedom, equality, and neutrality or toleration. Those requirements will be assessed on an individualised basis – manifesting, in practical terms, as legally guaranteed individual rights, for example. The rationality component will manifest as a claim that the theory is rationally

[26] Compare Rawls, 1999a: 74-75.

[27] Note that the formulation of a scope-based conception leaves open the content of 'freedom' (and 'equality' and 'neutrality'). We can generate a range of scope-based conceptions of liberalism by selecting specific conceptions to fill in the various component concepts. Thus, what we're really talking about here is a class of scope-based conceptions of liberalism, rather than a single conception. See Waldron, 1987: 130-134; Gaus, 2000: 77-157.

defensible – it will not rely on an appeal to religious or mystical authority, say. Insofar as the theory rejects cultural relativism, it can claim universal application. Thus, a scope-based conception is also a viable conception of liberalism. As an example of a scope-based conception, take the liberalism described by John Stuart Mill.[28]

We now have two alternative conceptions of liberalism. Evidently, the two might come apart. It is possible that citizens might (hypothetically) choose political institutions under which they have freedom of a relatively narrow scope, and it is possible that institutions under which citizens have freedom of a relatively wide scope could be imposed without reference to the authorisation of the citizens. And so, a particular set of political institutions might be liberal according to the authority-based conception, while not being liberal according to a scope-based conception, and *vice versa*.

There is obviously much to be said in favour of a scope-based conception of liberalism. It is straightforward and intuitive. Further, consider two objections that could be offered against the authority-based conception. First, isn't it strange to claim that a highly restrictive, authoritarian political system is liberal, simply because a bunch of citizens happen to agree to the system? Imagine, for instance, that a community of extremely devout people agreed to be governed according to a strict interpretation of a religious text. Surely the resulting political institutions would not be liberal. Second, take political institutions of the kind proposed by Mill. Suppose these institutions were imposed on the aforementioned religious community by a liberal dictator.[29] Are we to conclude that governance in accordance with the harm principle would not be liberal? That looks like an odd claim. So, it seems the authority-based conception tells us that political institutions are liberal when they clearly aren't, and that they aren't liberal when they clearly are. Shouldn't we then

[28] 1974.

[29] If you think that democracy is an essential component of liberalism, and thus that the idea of a liberal dictator is incoherent, suppose that the political system imposed is democratic, but constrained by a bill of rights with robust judicial review. Such a system is plausibly liberal, but might be imposed against the wishes of the citizenry, who would not be willing to endorse the freedom to engage in blasphemy, or other forms of sinful behaviour protected by the bill of rights.

reject the authority-based conception in favour of a scope-based conception?

In spite of these concerns, we should retain an authority-based understanding of liberalism. In order to defuse the two objections, note that they implicitly take the authority-based conception to depend on actual choice, rather than hypothetical choice. Unless we have some reason to think that ideally rational citizens, under conditions in which they are free and equal, would all have sufficient reason to choose theocratic political institutions, the circumstances imagined by the first objection could never arise. Similarly, under what conditions could free, equal, and rational citizens lack sufficient reason to choose Mill's institutions? It is true that, at this point in the argument, what it means for agents to have sufficient reason to choose a set of institutions remains underspecified. I take up that question later, when I discuss rational choice in greater detail. But the general shape of a reply to the two objections should still be clear at this point: The authority-based conception relies on hypothetical choices made by agents who have been idealised in certain important respects; insofar as those agents would have sufficient reason to choose all, and only, those political institutions which would be endorsed as liberal by plausible scope-based conceptions, the two objections lose their force.

One might be dissatisfied with this reply for one of two reasons. First, if the authority-based conception idealises its choosing agents in such a way that it does not differ from a scope-based conception in the institutions it identifies as liberal, what advantage does the authority-based conception have over the more straightforward scope-based conception? Call this the *parsimony* worry. Second, on the contrary, what if specifying agents as free, equal, and rational is not sufficient to limit the authority-based conception's definition of liberal political institutions to those which would be described as liberal by whatever we think is the most plausible scope-based conception? If that is the case, don't the two objections return with full force? Call this the *inadequacy* worry.

Take the parsimony worry first, and assume that the authority-based conception and a plausible scope-based conception identify the same range of political institutions

as liberal (thus, assume the inadequacy worry is unfounded). Nevertheless, the authority-based conception has advantages over the scope-based conception. The authority-based conception gives us a method of systematically organising the more primitive concepts (freedom, equality, etc.) from which liberalism is constructed. Each of these concepts plays a well-defined role in the hypothetical choice situation upon which the conception relies. On the authority-based conception, political institutions of a particular scope follow deductively from our understandings of the more primitive concepts. By giving us a *theory* of the relationship between liberalism's components, the authority-based conception renders our set of beliefs about politics more coherent and, contrary to initial appearances, more parsimonious. We can use this theory to create logical extensions of our beliefs, covering novel cases. A scope-based conception, by contrast, will often rely on *ad hoc* appeals to intuitions about the proper way to integrate and balance the primitive concepts.[30] It will typically lack the theoretical resources to explain why the primitive concepts are arranged in one way rather than another.[31]

Quite apart from these methodological considerations, the authority-based conception has a substantive advantage over a scope-based conception. It allows us to conceive of liberal political institutions as institutions under which people are not only free in the everyday sense characterised by what I have called 'wide scope', but are also free in the sense that they are living under political arrangements which they would have chosen, had they (or ideally rational versions of them) had the opportunity. On the assumption that political society cannot literally be a voluntary

[30] The authority-based conception does not, of course, eliminate the need for an appeal to intuition. We will be relying on intuition for our understandings of the more primitive concepts. However, the point is to relocate the appeal to a place in the argument where our intuitions seem likely to be more secure. See Rawls's discussion of intuitionism (1999a: 30-40). Contractarians, in particular, will note his claim that "dependence on intuition can be reduced by posing more limited questions and by substituting prudential for moral judgement." (*Ibid*.: 39.) Compare also Rawls, 1985: 227, 229; Gauthier, 1986: 269.

[31] Rawls, 1999a: 36-37. Not all scope-based conceptions will have this feature. A scope-based conception based on utilitarianism could use the utilitarian principle to adjudicate systematically between different configurations of the primitive concepts (*ibid*.: 36). However, to the extent that a scope-based conception does this it will, first, become less straightforward, inasmuch as the conceptual structure of the conception will have become more elaborate. Second, it will expose itself to versions of the inadequacy worry. In both respects, the scope-based conception would then lose its supposed advantages over the authority-based conception.

association, we respect individuals' autonomy to the greatest extent possible by paying attention to the requirements of hypothetical choice.[32] The authority-based conception thus captures an additional, deeper sense of freedom which is constitutive of liberalism.

Thus, the parsimony worry shouldn't concern us. The authority-based conception has both methodological and substantive advantages over a scope-based conception, even on the assumption that both conceptions identify the same range of institutions as liberal.

Next, consider the inadequacy worry. Suppose that the most plausible way of characterising free, equal, and rational agents leads the authority-based conception to differ from a plausible scope-based conception in the range of institutions it identifies as liberal. If this is the case, then the authority-based conception is open to at least one of the two objections I described above. Either it will identify as liberal some institutions we intuitively regard as illiberal, or it will identify as illiberal some institutions we intuitively regard as liberal (or both). In any event, the worry is that the idealisation of agents built into the authority-based conception does not adequately close the gap between the conception and our pre-theoretic understanding of liberalism.

It is difficult to give a simple response to this concern. It is reasonable to insist that there are limits beyond which the concept of liberalism cannot plausibly be stretched.[33] On the other hand, we shouldn't be afraid to be at least moderately revisionist about our understanding of liberal political institutions, if we have good reason to be. As noted above, we may have such a reason: The authority-based conception incorporates a radical sense of freedom as autonomous choice which scope-based conceptions neglect. The extent to which we should be willing to revise our pre-theoretic understanding of what constitutes a liberal set of political institutions will track the degree to which we think it is important that our conception

[32] Rawls, 1999a: 12. See also the discussion of free choice in §4.2, *infra*.
[33] Gauthier, 1991: 29.

of liberalism reflects this sense of freedom.

The authority-based conception's radical sense of freedom has two aspects which make it attractive as a component of a conception of liberalism. First, in understanding the state as something that is chosen (albeit hypothetically) by citizens, it makes individuals conceptually prior to the state in a fundamental way. On this view, an individual is not simply a token for her society, family, culture, or nation; she is not merely a tool or component part of some larger entity. She has an existence independent of these entanglements. Second, this sense of freedom allows us to distinguish between limits on the scope of our ability to act which we have chosen and limits which we have not chosen. It allows us to say that freedom includes the ability to make commitments – among other things, to make promises, to enter into contracts, and thus to *exercise* one's freedom in a way that limits one's available range of action.[34] There is a difference between me choosing to foreclose the possibility of *B* in order to realise *A*, and you simply eliminating the possibility of *B* and leaving me with *A*. But this difference cannot be expressed in terms of the scope of my ability to act, because the two cases do not differ in that respect. Similarly, we might plausibly want to distinguish between me taking your car, *simpliciter*, and me taking your car because you have entered into a contract with me to sell the car. In both cases, you will lack the car once I have taken it, but a liberal conception of freedom should be able to distinguish between a theft and a contract (or a gift).[35]

Taking these two aspects together, the authority-based conception understands citizens themselves to have the authority to make and remake states and societies as they choose.[36] In choosing the constraints – laws, political institutions – under which she lives, a citizen is free. She is acting in accordance with her rational will.[37] On

[34] Waldron, 1987: 133-134; Stilz, 2009: 30.
[35] Waldron, 1987: 133-134.
[36] At this point, we are not yet committed to any kind of *democratic* authority. Something like majority rule is significantly further down the chain of argument than I have travelled here.
[37] The most famous statement of this view is Rousseau's in *The Social Contract*, Book I, Chapter 8 (1968: 65). Compare Locke, *The Second Treatise of Government*, §57 (2005: 40-41). More controversially, see Mill, *On Liberty*, Chapter V (1974: 166), and the discussion of this passage by Gaus (2011: 33-36).

this view, whether or not a set of political institutions is liberal depends on the ability of citizens to endorse the framework which constrains their everyday choices (what Rawls has called the 'basic structure'), rather than simply on the scope to choose which citizens have within that framework.[38] A scope-based conception of liberalism goes wrong because it focuses on the latter rather than the former. In an anarchic society, for example, the scope of citizens' choices might be very wide (and might be equally wide, etc.), but that would not make such a society liberal. Liberalism is not just a diluted or compromised form of anarchism, and the authority-based conception gives us a way of articulating what is distinctive about liberalism's understanding of the relationship between freedom and law.[39]

Thus, we shouldn't worry if adopting the authority-based conception results in a moderate revision of our pre-theoretic understanding of liberalism. Our pre-theoretic understanding is not entitled to absolute deference, and we have good reason to revise it if it would not reflect the choice of free, equal, and rational citizens. Insofar as the revisions prompted by the authority-based conception are not excessive, then, the inadequacy worry should not trouble us.

Despite this rejection of the inadequacy worry, I don't deny that when asked to choose a set of political institutions to govern their interactions, free, equal, and rational citizens likely would select a framework which allows them freedom of a wide scope in their everyday lives. That they would likely do so reflects the value people typically attach to having certain kinds of options, and the typical diversity of citizens' tastes, ambitions, and moral commitments. These are, however, contingent features of the world. In their absence, people would choose different frameworks within which to live, and, on my view, liberalism would then recommend political institutions with a narrower scope. My claim is thus that political institutions with wide scope are not strictly constitutive of liberalism, as a scope-based conception holds. Rather, they are the result of applying liberalism in

[38] 2005: 265-266.
[39] Waldron, 1987: 134; Korsgaard, 1996: 92-97, 105-107; Ryan, 2012: 24; Moehler, 2018: 134-135.

a world with particular contingent features (albeit, a world we are unlikely to be leaving any time soon). The constitutive core of liberalism is instead the requirement that political institutions be justifiable *to* free and equal agents.

It has not been my aim here to give a definitive account of liberalism, or to adjudicate between scope-based and authority-based conceptions with any finality. Liberalism is a contested concept which admits a variety of reasonable interpretations.[40] Nonetheless, the authority-based conception has enjoyed significant support among contemporary theorists. Because it is based on hypothetical and not actual choice, it can answer the naïve objections I have considered. Moreover, it has both methodological and substantive advantages over the scope-based conception which are sufficient to overcome the more sophisticated parsimony and inadequacy worries I have discussed. Thus, we can take the authority-based conception as a working understanding of liberalism.

§4.2 The authority-based conception and contractarianism

In my remaining chapters, I will develop a model in which agents hypothetically choose political institutions to regulate their interaction with one another. The agents will be characterised as rational, and will make their choices on the basis of their prudential reasons. The model is thus a contractarian model.[41] I will argue that each agent in the model has decisive prudential reasons to agree to coordinate her actions with others in establishing and maintaining a set of political institutions under which the goals of each agent are advanced to an equal degree.

The political institutions which emerge from the model will be liberal political institutions because, first, the model incorporates all of the necessary components of an authority-based conception of liberalism; and, second, the character of the institutions selected will not diverge excessively from our pre-theoretic

[40] Quong, 2011: 11; Rawls, 2005: 6.
[41] §1.1.

understanding of liberalism. Demonstrating this second claim will not be possible until the argument of the later chapters is complete, but I can lay the groundwork for the first claim now.

The agents in the model will rationally choose political institutions on the basis of their reasons; each individual agent will be shown to have sufficient reason to choose the institutions; and the model will (within the limits described by the circumstances of justice) abstract from any particular cultural or temporal context.[42] Thus, the model satisfies the rationality, individuality, and universalism requirements of the authority-based conception.

The model will also characterise the agents as free and equal, though whether or not it will do so to the satisfaction of a contemporary liberal philosopher is an open question. The agents in the model will be able to choose the manner of their political coordination, or to reject coordination entirely in favour of conflict.[43] They will not, however, be restrained from coercing or threatening one another in order to secure more advantageous terms of coordination.[44] Thus, the sense in which they *freely* choose the institutions which emerge from the model is attenuated.

One might think that to grant agents the austere freedom of the highwayman's victim – the ability to choose between some terrible fate and compliance with a threat – is to grant them no freedom at all.[45] But this is too quick. In all of the classical social contract theories, citizens are driven into political society by the threat of the state of nature, whether that threat comes in the form of a vulnerability to natural dangers (wild animals, starvation, exposure) or artificial dangers (the predation of other

[42] Compare Moehler, 2018: 28, 164.
[43] See Binmore (1994: 209) for a discussion of whether agents who can only choose rationally can choose freely.
[44] This is, of course, an explicit departure from Rawls's approach (see, for example, 1985: 235).
[45] Compare Ripstein, 1987: 116, 133-134; Binmore, 1994: 72; Rossi, 2010: 30-38; Ryan, 2012: 10-12. Gauthier attempts to avoid this problem by requiring that agreements between agents are not based on coercion (1986: 190-232; also Danielson, 1991; Hubin and Lambeth, 1991; Narveson, 1991; Binmore, 1994: 12-18). As I will explain in §7.5, I do not find Gauthier's argument convincing, and so contractarianism seems to me to be stuck with the austere conception of free choice I defend here.

people).[46] The mere existence of threats, even existential threats, in the state of nature doesn't make our choice to enter political society unfree on the classical accounts.[47]

Perhaps this is a reason to reject or revise the classical accounts. Note, however, that this is easier said than done. Suppose we think that, given a choice between *A* and a set of other options, a choice of *A* can be genuinely free only if there is at least one option, *B*, which is an acceptable alternative to *A*. The problem then is that, if *B* really is an acceptable alternative to *A*, why wouldn't at least some agents choose it? If *A* is the choice to enter political society, what is *B* supposed to be? If we suppose that entering political society carries significant costs – the requirement to obey laws, to pay taxes – rational agents will only enter if doing so provides them with significant compensating benefits which they cannot secure in the state of nature. Such benefits are typically thought to include protection, the prospect of capital accumulation, and technological progress. These benefits, however, are so important to our well-being that they are precisely the kinds of things whose absence would render any alternative unacceptable. If we can *only* secure them by entering political society, then there can *never* be an acceptable alternative to entering political society, and thus the choice to enter political society can never be free in the robust sense I am presently considering. If, on the other hand, we can secure these important benefits *without* entering political society, we would have no reason to take on the significant costs which come with citizenship, and so the revised social contract theories could never justify the state. Insisting on the robust sense of free choice is therefore self-defeating for social contract theories.

It may be thought that Rawls offers us a way out of this dilemma. Rather than trying to come up with a plausible motivation for agents to enter political society when, by virtue of being free, they have an acceptable alternative, Rawls simply stipulates that

[46] Hobbes, in *Leviathan*, Chapters XIII, XIV, and XVII (1996: 86-100, 117-121); Locke, in *The Second Treatise of Government*, §13, Chapter III, and Chapter IX (2005: 22, 24-26, 71-74); Rousseau, in *The Social Contract*, Book I, Chapter 6 (1968: 59; compare 1984: 109-125).

[47] Similarly, if we think the fact I am subject to a threat renders me unfree to choose in the relevant sense, it is hard to see why we would be precious about the distinction between threats posed by other agents and dangers posed by nature which other agents exploit when they offer to deliver me from the state of nature (see §3.2.2, particularly n43).

the agents are motivated to enter.[48] In the original position, then, the agents' choice is what kind of society they will enter, not whether they will enter a society. Could the same move work here?

I don't believe so. Recall that the purpose of the constructive exercise here is to identify a set of political institutions as liberal. In terms of the content of the set of institutions selected, what is the difference between being forced into agreement by natural or artificial threats and being forced into agreement because the alternative to agreement has been excluded by theoretical fiat? If there is no difference, then the Rawlsian move is extensionally equivalent to adopting the austere conception of free choice.

Perhaps this is too quick. One possible difference might be that, while theoretical constraints on one's options affect all agents equally, some agents might be more vulnerable to threats (or more capable of making threats) than others. The real worry, then, is about the content of the social contract that will result if coercion and exploitation are allowed to influence the terms of the agreement. Surely, the objection will go, if the strong are allowed to threaten or exploit the weak, they will end up with the bulk of the benefits of political coordination. No such agreement could be called liberal.

As far as I can see, however, this is not best understood as a worry about freedom; rather, it is a worry about equality. For the purpose of developing the contractarian model, I will for the most part simply assume that the agents in the model are equally powerful. (Chapter 8 will then argue for this assumption.) With the assumption of equal power in place, it isn't clear why there would be any reason to think that the agents, when they come to some agreement, have not done so freely, despite the fact that there has been no formal restriction on them threatening one another. If

[48] This is not quite the whole story. Rawlsian agents are motivated to act justly, and they acknowledge a natural duty to establish just institutions (1999a: 293-294; see also §3.2.2, *supra*). Insofar as, in the absence of just institutions, justice would not be realised, Rawlsian agents are motivated to enter political society by the possibility of an unjust state of affairs obtaining. Nonetheless, the existence of this possibility does not make their choice of political institutions unfree. See also Thoma, 2015: 3336.

threats don't have a disparate impact – if there is equality of 'threat advantage' – we're led back to the conclusion that there is no difference between making the Rawlsian stipulation and adopting the austere conception of free choice.[49] Thus, the way in which the contractarian model characterises its agents will satisfy both the freedom component and (granting the argument of Chapter 8) the equality component of an authority-based conception of liberalism.

This leaves the neutrality component. The manner in which the contractarian model incorporates a neutrality restriction is two-fold.[50] Its first aspect tracks a version of *justificatory* neutrality; its second tracks *consequential* neutrality.[51] The political institutions prescribed by the model will be justificatorily neutral because they do not rely for their justification on any conception of the good (or, for that matter, the right) not accepted by the agents to whom the prescription is addressed. Since the prescriptions of the model are based on each agent's prudential reasons, they will necessarily satisfy justificatory neutrality (regardless of their content).

Justificatory neutrality is an extension of the idea of conceiving of the agents as free, because it means that agents are not asked to act for reasons which are alien to them.[52] By contrast, that the institutions prescribed by the model are consequentially neutral is an implication of assuming equal power – consequential neutrality is thus an extension of conceiving of the agents as equal. Consequential neutrality is the requirement that the actions of the state not have the consequence of advancing one conception of the good (held by a citizen) over other conceptions of the good (held by other citizens).[53] Consequential neutrality has generally been seen as a wildly

[49] See Rawls, 1999a: 116, 122.
[50] Compare Zellentin, 2012: 49-51.
[51] Kymlicka, 1989: 883-886; Sher, 1997: 3-4. For further references, see n17, *supra*.
[52] I am here understanding justificatory neutrality as a constraint on the form of political institutions – on "constitutional essentials and matters of basic justice" – rather than as a constraint on the actions certain political institutions take in their day to day operations (Rawls, 2005: 442; compare Quong, 2011: 273-287). Nevertheless, this does not rule out also applying justificatory neutrality to the day to day actions of political institutions. Whether or not the actions of political institutions are so constrained is a question about the substantive content of the agreement to which the agents in the contractarian model come (see the discussion of consequential neutrality in the text, and §7.8, *infra*).
[53] The parentheses serve to emphasise that the model doesn't care if political institutions fail to be consequentially neutral with respect to any conception of the good which is not actually held by some citizen (Raz, 1986: 110-112).

implausible candidate for a neutrality principle, primarily because it is thought that the state cannot possibly comply with its requirements – basic and essential state functions such as maintaining law and order by establishing a police force will, it is said, inevitably have the effect of promoting conceptions of the good which prize discipline over those which prize spontaneity.[54]

Despite these concerns, I will endorse a version of consequential neutrality in setting out the contractarian model. As I will explain, the agents in the model are assumed to be aiming to advance their ends to the greatest extent possible. Since it won't be possible for them all to maximally advance their ends at the same time, they will compromise and select political institutions which advance each agent's ends to some degree. Because they are assumed to have equal bargaining power, the point at which they compromise will be the point at which each agent's ends are advanced to an equal degree. Insofar as an agent's set of ends is equivalent to her conception of the good, the model will therefore recommend political institutions under which agents' conceptions of the good are advanced to an equal degree – thus, it will require consequential neutrality.

The concerns about consequential neutrality raised in the literature are perfectly sensible if one is looking for a principle to guide state policy on a day to day basis. On that level, consequential neutrality is impractical. These kinds of concerns should not trouble the contractarian argument, however, for two reasons. First, they point to difficulties associated with implementation, but the contractarian argument is aimed at a higher level of abstraction. My purpose is to identify the principles to which a set of political institutions (prudentially-)ought to conform, given certain idealisations of the world, rather than to describe procedures for implementing those principles. Applying the theory to a particular real-world context obviously raises many complications which the theoretical argument neglects.[55] Second, when we do move to the level of implementation, it is plausible to suppose that the agents

[54] *Ibid*.: 110-124; Sher, 1997: 3-4; Quong, 2011: 18.
[55] See the remarks on models in the Interface. Compare Dworkin, 1981a: 196-197; Dworkin, 1981b: 292.

will seek to establish political institutions which equally distribute the *means* of advancing ends, rather than seeking to achieve equal advancement of ends directly through the actions of political institutions. Proceeding in this way will allow the agents finer control over the precise timing and manner of the advancement of their ends, and will allow them to respond to changes in their preferences more quickly and effectively (complications neglected at the level of principles). Distributing the means of advancing ends equally – in the manner of Rawlsian primary goods, for example – does not raise the same kind of practical difficulties as attempting to implement consequential neutrality directly.[56]

It should be clear, then, that a contractarian model can in principle qualify as an authority-based conception of liberalism. In Chapters 5 and 6, I will develop such a model. In Chapters 7 and 8, I will argue that the model I have developed (i) meets the criteria for a liberal theory described in this chapter, and (ii) does so without any compromise in its essential contractarianism (that is, without relying on any moral foundations). This will allow me to draw two interesting conclusions. First, that in the conditions of contemporary societies, contractarianism is committed to genuinely liberal political institutions. And second, that the problem of insufficient justness – raised in Chapter 3 – can be resolved.

[56] §7.8.

Chapter 5

A contractarian model of political society is composed of four elements: rational agents, their preferences, their circumstances, and a prescription for action. A prescription is addressed to a rational agent, on the basis that it is the most effective means of realising his most preferred state of the world, given the facts about the world in which he is acting.

In this chapter, I discuss the first three elements of the model. In §5.1, I explain what preferences are, and how an agent's beliefs and dispositions for action must be related to his preferences in order for the agent to qualify as rational. In §5.2, I state the conditions which an agent's set of preferences must satisfy in order for the agent to qualify as rational. In §5.3 and §5.4, I defend two of these conditions – completeness and transitivity – against prominent objections. In §5.5, I state and defend some claims about the way my contractarian model represents the world in which its agents operate. These three elements – the agents, their preferences, and their circumstances – constitute the precontractual conditions from which the fourth element, the contractarian prescription for political society, will be derived.

§5.1 Preferences and rationality

I begin with agents, and with the idea that agents have preferences. To say that an agent has a (strict) *preference* for a state of the world, S_1, over another state of the world, S_2, is to say that the agent believes herself to have greater reason to act in such a way as to realise S_1 than S_2.[1] We can say an agent is *indifferent* between S_1

[1] Compare Hansson, 1993: 482; Sobel, 1997: 55-56, 67-71; Gray, 2000: 45; Rabinowicz, 2000: 124, 143-144. I begin with the idea of strict preference simply for ease of informal exposition. This should not be taken to imply that I identify strict preference formally as the primitive relation. Weak preference ($S_1\mathbf{R}S_2$) is usually defined as the primitive relation – in my schema, to say that an agent has a weak preference for S_1 over S_2 would mean that the agent believes herself to have a reason to realise

and S_2 if and only if she believes herself to have a reason to realise S_1 which is at least as great as the reason she believes she has to realise S_2, and also believes that she has a reason to realise S_2 which is at least as great as the reason she believes she has to realise S_1.[2]

Given the kind of creatures human beings are, their preferences will typically include preferences for having basic things like food and shelter rather than lacking them, and, more generally, for the absence (rather than experience) of pain and the experience (rather than absence) of pleasure. But their preferences may also include more sophisticated things: a preference for playing golf rather than tennis, or listening to rap instead of reggae, or engaging in family life rather than in religious worship or in scientific research. Some preferences are shared by (nearly) all human beings; other preferences will be unique to individual agents.

The conception of 'preference' I employ is a purely formal one.[3] I do not assume that agents are selfish, self-interested, or mutually disinterested (nor do I assume the contrary). Any particular agent will almost certainly have preferences relating to her own well-being, but there is nothing to stop her also having preferences relating to the well-being of others (whether she wishes those others good or ill), or preferring the realisation of her other-regarding ends over her self-regarding ends.[4] Similarly, an agent in the model may have preferences relating to the ways in which society is organised, to the realisation of certain conceptions of the good, or to realising a conception of justice.

S_1 which is at least as great as the reason she believes she has to realise S_2 (strict preference is then denoted $S_1\mathbf{P}S_2$, where $S_1\mathbf{P}S_2 \Leftrightarrow S_1\mathbf{R}S_2 \wedge \neg(S_2\mathbf{R}S_1)$). In the text, I have described preferences in terms of beliefs agents have about their objective reasons (see §2.3.1 n29, *supra*). Defining preferences in terms of beliefs about reasons for action is not standard, but it allows me to accommodate moral beliefs in the model quite straightforwardly. Compare Smith's discussion of the distinction between valuing and desiring, and his related analysis of normative reasons (1994: 130-181).

[2] In the standard notation, $S_1\mathbf{I}S_2$, where $S_1\mathbf{I}S_2 \Leftrightarrow S_1\mathbf{R}S_2 \wedge S_2\mathbf{R}S_1$ (see n1). I thank David Birks for helpful discussion of these points. See also Hansson, 2001: 23-26.

[3] §1.3 n16; Moehler, 2018: 102. Compare Hobbes's distinction, in The Introduction to *Leviathan*, between men's passions and the objects of those passions (1996: 10).

[4] Thus, the model allows agents to have 'tuistic' preferences. See Gauthier, 1986: 100-104; Ripstein, 1987: 129-130; Hubin and Lambeth, 1991: 122-124; Binmore, 1994: 25, 44 n47, 51, 79, 286, 288; Laver, 1997: 27-28; Rawls, 1999a: 12-13, 123-124, 128-129; Moehler, 2018: 102. For a discussion of the 'non-tuism' assumption in Gauthier's work, see Hubin and Lambeth (*op. cit.*), Morris (1991), Vallentyne (1991b), and Moehler (2018: 53, 102 n18).

This formalism allows us to dismiss quickly one common concern about models based on rational choice. It is sometimes objected that the *homo economicus* inhabiting such models is an impoverished creature, solely and selfishly interested in his own material well-being, to the exclusion of any feelings of solidarity, or love, or sympathy, or of any sense of moral duty. Given that this is not an accurate representation of real people, the worry is that any prescription based on the representation will be otiose.[5] But this objection rests on a simple mistake. The model does not assume that its agents lack any part of the full range of commitments which real people typically have. In particular, because it is intended to be fully compatible with moral realism, the model allows that agents may hold that their preferences over moral ends are not merely tastes but express fidelity to objective moral truths (as I do).[6] What it requires is simply that an agent, *J*, treat moral beliefs held by some other agent, *K*, which *J* believes to be false, as preferences held by *K*. In this way, the model can accurately reflect the range of moral commitments that people typically have, without the theory itself endorsing any of those commitments.[7]

In saying that agents are *rational*, I mean that their intentional actions stand in a particular kind of relationship with their preferences. To say that an agent is rational is to say, among other things, that she is motivated to act on the beliefs which comprise her preferences, all else equal.[8] So, given the choice between an action,

[5] Binmore, 1994: 18-28.
[6] §1.5.
[7] Compare Rawls, 1999a: 12, 123-130; Vanderschraaf, 2019: 5-7.
[8] Compare Broome (2013: 2), who says that "rationality requires people to intend to do what they believe they ought to do, and that it requires them to be disposed to do so..." In this claim, we can observe the three components of the relationship between preferences and rationality I want to highlight: (i) preferences are beliefs about the reasons one has to act; (ii) rationality requires that preferences and decisions about how to act be consistent; and (iii) absent special circumstances (external constraint, say), rationality requires that one's actual actions and one's decisions to act be consistent. (See also *ibid*.: 88-90; Broome, 2005a: 323.) One might say, by way of objection, that one can have preferences over states of the world without it following that one believes that one has (even a *pro tanto*) reason to realise those states. I do not wish to take a position on that question; I will simply note that I am using the term 'preference' more narrowly, to refer to beliefs about reasons for action, rather than purely evaluative beliefs, say. I should also note that Broome distinguishes between 'ought' and 'reasons', defining the latter in terms of the former, whereas I am using the two ideas interchangeably. By the unqualified 'ought', Broome means what I have called '(all-things-considered-)ought' (2013: 25-26, 63-64; and §1.3 n19, *supra*). It also bears mentioning that my

A_1, which will realise S_1, and an action, A_2, which will realise S_2; and given that S_1 is preferred by an agent to S_2; then, the agent is rational only if she decides to do A_1 rather than A_2.[9]

§5.2 Formal constraints on preference sets

One component of the rationality of an agent consists in the relationship between the agent's preferences and her intentional actions. A second component is concerned with the nature of the agent's preference set. In saying that an agent is rational, I will mean that her preference set has certain properties. First, rational agents have preferences over all possible states of the world of which they are aware. So, given two states of the world, S_i and S_j, an agent will either prefer S_i to S_j; or will prefer S_j to S_i; or will be indifferent between S_i and S_j. That is, the agents' preferences are *complete*. Second, rational agents have preferences which are *transitive*, so that if an agent prefers S_i to S_j, and also prefers S_j to S_k, then she necessarily prefers S_i to S_k; similarly, if she is indifferent between S_i and S_j, and also indifferent between S_j and S_k, she is necessarily indifferent between S_i and S_k.[10]

Given completeness and transitivity, we can produce for each agent an ordinal

---

position here goes beyond what contemporary revealed preference theory would claim. It would make the more modest claim that rational agents behave *as if* they have the beliefs and motivations I have ascribed to them, without making the claim that the agents actually have those beliefs and motivations (Binmore, 1994: 50-51, 104-106, 180-181; Bar-Hillel and Margalit, 1988: 141-143; Gauthier, 2013: 602).

[9] We might think that it is possible for an agent to prefer S_1 to S_2, but believe herself to be morally required to realise S_2, and act so as to realise S_2, and further that this would not be irrational. To the extent this is plausible, it relies on an understanding of preference that differs from the one to which I have stipulated (more or less, it relies on an understanding of preferences as something distinct from morality). There are contexts in which it is useful to be able to distinguish between what we might term 'mere preferences' and moral commitments. In such cases, we can understand a mere preference as an agent believing that she has a greater prudential reason to realise S_1 than S_2, and a moral commitment as an agent believing that she has a greater moral reason to realise S_2 than S_1. But, for the purpose of building the contractarian model, I want to be able to discuss a ranked ordering of all of the reasons for action (prudential and moral) which an agent believes apply to her. When I refer to an agent's set of preferences, or her preference set, it is this more inclusive understanding I will have in mind. Compare Arneson, 1990a: 161-164.

[10] Nash, 1950: 156; Gauthier, 1986: 38-42; McClennen, 1990: 20-21; Binmore, 1994: 104-105; Hansson and Grüne-Yanoff, 2017: §§2.1-2.3.

ranking of all available states of the world. However, what we cannot yet do is say *how much* a particular agent prefers S_i to S_j. In developing the contractarian model, it will be helpful to be able to produce such a cardinal measure.

It will be possible to construct such a measure if the agents' preferences satisfy two further conditions.[11] First, I assume that the preferences are *monotonic*.[12] Consider two bundles of goods, *X* and *Y*. *X* consists of goods $\{x_1, x_2, x_3 \ldots x_n\}$. *Y* consists of goods $\{y_1, y_2, y_3 \ldots y_n\}$. An agent's preferences are monotonic if and only if *X* is preferred to *Y* when, for at least one x_i and y_i pair, x_i is preferred to y_i, and for all other x_j and y_j pairs, x_j is at least as good as y_j. To illustrate, suppose I'm comparing two sports cars, a Ferrari and a Lamborghini. Suppose the cars cost the same, and that each is just as fast as the other. However, the Ferrari only comes in red, while the Lamborghini only comes in yellow. The monotonicity condition says that, in order to be rational, my preference over the cars must depend on whether I prefer red or yellow, and must correlate positively with my colour preference, because the other relevant dimensions (cost and speed) don't distinguish the options. To put that another way, the monotonicity condition requires that the various dimensions upon which I am assessing my options are genuinely independent of one another. Thus, it can't be the case that I prefer higher speed in cars that are red, but prefer lower speed in cars that are yellow.

Importantly, the monotonicity condition applies to possible lotteries over which agents can have preferences. By way of example, consider two: in the first lottery, I will win *X* with a probability of *p*, and *Z* with a probability of $(1 - p)$; in the second lottery, I will win *Y* with a probability of *p*, and *Z* with a probability of $(1 - p)$.[13] These lotteries are like bundles – the first bundle contains a chance of getting *X* and a chance of getting *Z*; the second bundle contains a chance of getting *Y* and a chance of getting *Z*. Since the chance of getting *Z* is the same in each bundle, my preference

[11] I follow Gauthier (1986: 42-46) here; see also Nash (1950: 156).
[12] What Gauthier calls monotonicity is often described as the *independence* condition or the *sure-thing* principle. It should not be confused with the *independence of irrelevant alternatives* condition discussed in Chapter 6. See McClennen, 1990: 44-59, 67-80; Broome, 1995: 24-29, 60-89, 94-95.
[13] Note that the value of *p* is constant across these lotteries.

over the bundles must be determined by my preference over X and Y. If I prefer X to Y, by monotonicity it follows that I must prefer the first lottery to the second.

The second further condition is *continuity*. Suppose an agent (with transitive preferences) prefers S_i to S_j, and S_j to S_k. Her preferences are described as continuous if there is some value p such that she is indifferent between S_j and $(pS_i + (1 - p)S_k)$. We can understand this latter expression as a lottery, with p as a probability, and S_i and S_k as possible prizes. Continuity implies that there is such a lottery, in which the agent is indifferent between the certainty of receiving her middle-ranked option, S_j, and being offered a gamble with a probability p of winning her more preferred option, S_i, and a probability $(1 - p)$ of winning her less-preferred option, S_k.

With these two additional conditions in place, we can transform ordinal rankings of options into cardinal rankings.[14] First, we take some highest ranked option, $\overline{S}$, and assign it a value of 1; then, we take a lowest ranked option, $\underline{S}$, and assign it a value of 0. Finally, we take any pair of options S_i and S_j, and find the lotteries over $\overline{S}$ and $\underline{S}$ for which an agent is indifferent between the lottery and each of S_i and S_j. Since the values of $\overline{S}$ and $\underline{S}$ are 1 and 0 respectively, the value of the lottery $(p\overline{S} + (1 - p)\underline{S})$ over them is simply p. If an agent is indifferent between the respective lotteries and S_i and S_j at different values of p, we will be able to say with precision how much she prefers S_i to S_j by comparing the relevant p-values.[15] In principle, we can do this for all of the options available to the agent. Thus, we have a cardinal scale for the agent's preferences which can measure the intensity of her preferences as well as their rank order. This scale is called a von Neumann-Morgenstern (vNM) utility function.[16] Being able to give such a measure will be useful later in the argument, because I will want to be able to say how much each agent is gaining or losing from a proposed set of political institutions.

A full defence of the four formal conditions I have described isn't possible within the

[14] Again, I follow Gauthier here (1986: 44).
[15] Note that p is now acting as a variable, rather than a constant. Compare n13.
[16] Binmore, 1994: 269-279, 282; Peterson, 2009: 94-103, 302-307.

scope of my discussion. They are not uncontroversial, but they are standard assumptions in rational choice models.[17] In lieu of a full defence, I will comment on objections to the two more fundamental conditions – completeness (§5.3) and transitivity (§5.4). I do not mean to imply that monotonicity and continuity are universally accepted. Monotonicity, in particular, is the target of two famous objections – the so-called Allais and Elsberg paradoxes.[18] I refrain from further discussion of these two conditions for three reasons. First, such a discussion would become increasingly technical, and would take me increasingly far from my topic into debates in decision theory. Second, the examples used to support Allais- and Elsberg-type objections are quite exotic, and we have reason to be wary of appeals to intuition drawn from such examples. Finally, monotonicity and continuity have simply not (yet) been much discussed or challenged by political philosophers.

In contrast, completeness and transitivity have attracted considerable critical attention. The possibility of deriving determinate prescriptions from the contractarian model depends on rejecting this criticism. I also have one additional reason for discussing objections to completeness and transitivity in some detail. Part of the defence of the two relies on showing that we cannot plausibly think an agent rational if she fails to meet either of the conditions. In Chapter 8, my argument for agents in the contractarian model having equal bargaining power requires an extension of the standard account of rationality. I will argue there by analogy with the defence I give of completeness and transitivity here. Thus, §5.3 and §5.4 provide groundwork for the argument of Chapter 8.

§5.3 Objections to completeness

The first set of objections is aimed at the completeness condition. Isn't it implausible

[17] The classic discussion is given by von Neumann and Morgenstern (1947: 26-29). See also Nash, 1950: 156-157; Gauthier, 1986: 21-26; Kraus, 1993: 5-6; Binmore, 1994: 304; Laver, 1997: 18-25; Peterson, 2009: 164-171, 176-183; Moehler, 2018: 100-101.
[18] McClennen, 1990: 67-80, 91-95; Broome, 1995: 94-100, 107-115; Peterson, 2009: 78-83, 176-183.

to suggest that any actual person could have preferences over *all possible* states of the world of which she is aware? Actual people simply don't have that kind of mental processing power available to them.[19] Insofar as we want our contractarian model to generate prescriptions for people in the real world, this may seem problematic.

There are two different ways in which an agent can have incomplete preferences. On the one hand, there is an informal sense of 'incomplete'. I don't consciously have complete preferences over all possible meals I could order from my local takeaway, in the sense of being able to list them on request, because I haven't gone to the trouble of making systematic, pairwise comparisons between all of the possible combinations of options on the menu. In principle, however, there is nothing stopping me from doing so. While I could not, on demand, recite the various options in order of my preference, there is a sense in which I do actually *have* those preferences. If you presented me with any particular pair of options, I would be able to choose between them without any difficulty (or, alternatively, to express indifference). I would choose by consulting my preferences, and my capacity to do so implies that my preference over any pair of options does exist, even if I don't have access to it until I direct my attention appropriately.[20] As such, even if I cannot

[19] See the discussion of processing power in the Interface.

[20] This claim might be thought controversial – see Anand (1987: 191-192), and Hansson (1989: 3; 1993: 483). Hansson distinguishes between two different ways of understanding the extension of preference sets to include elements which are not initially accessible to actual people. He identifies the methodology I describe as 'the discovery account'. His preferred alternative is 'the development account'. On the development account, at least some of my preferences come into existence when I am presented with options I have not considered before, rather than already being in existence and being discovered by me through introspection. Hansson gives us two reasons to prefer the development account to the discovery account. First, he tells us that the development account fits better with a commitment to value pluralism (1989: 3; 1993: 483); second, the development account can accommodate the fact that making comparisons between options is often costly for agents – for example, it allows us to say that an agent is not irrational if she acts contrary to what her considered judgement would be because reaching a considered judgement ('discovering' her preferences) would have non-trivial costs which outweigh the (expected) benefits of engaging in (further) deliberation (1993: 483). Neither of these strikes me as a sufficiently persuasive reason to abandon the discovery account. I am willing to say that the circumstances of the agents in the contractarian model are idealised in such a way that the second problem does not apply (Gauthier, 2013: 603). I take these requirements of the discovery account to fall within the scope of the 'quite radical' idealisation I describe in the Interface (*supra*). With regard to the value pluralism problem, the discovery account does not require that agents understand themselves to be pursuing a single master value. What it requires is that they have a well-defined understanding of the ways in which the various values to which they are committed interact with one another. It may be argued that, to the extent that relations between the several values are well-defined, there is necessarily an implicit meta-value present, and so the model is actually a value monist one. If this is true, and if it is problematic, it seems to me that

articulate a complete set of preferences over all possible states of the world, it is possible to construct a complete preference profile for me by asking how I would choose between pairs of options if, hypothetically, I were presented with them.[21] I intend my conception of idealisation to include this possibility. Thus, this kind of (informal) incompleteness isn't a problem for the model.

On the other hand, there is a more formal sense in which preferences can be incomplete. Perhaps my preferences are incomplete, not because I haven't thought hard enough about my options, but instead because some of my options are *incommensurable*. The classic example here comes from Jean-Paul Sartre.[22] Suppose it is 1942 and France has been occupied by Nazi Germany. You are a French citizen who has an opportunity to leave France and join the Free French Forces. However, you live with your mother, who needs your care and support. Should you join the Free French or stay home to look after your mother? In cases such as this, you might plausibly say that you are genuinely torn between your options. No matter how much time we allow you to deliberate, you would be incapable of coming to a judgement that you have greater reason to pursue one option than the other. Nevertheless, you might plausibly deny that you are indifferent between the options. It does not accurately capture your state of mind to say that you are equally happy to pursue either option, or that it is of no consequence to you which option you pursue. Rather, it matters very much to you, but no matter which option you end up choosing, there is a sense in which you will regret your choice, and wish you had chosen differently. Each option is valuable in a different way, and the value of one is incommensurable with the value of the other. Since the completeness condition denies that agents can have pairs of options with respect to which neither a preference nor an indifference relation exists, the contractarian model fails to accurately capture some of our most important beliefs about values.

---

it can only be problematic as a result of some commitment to value incommensurability. I discuss incommensurability in the text, *infra* (see also Crowder, 2007: 125-126, 136-137).

[21] Compare Enoch's discussion of implicit beliefs (2015: 226-230).

[22] 1948: 35-38, 69-70. To better illustrate my point, I have taken some liberties with Sartre's example. See also Galston, 1999: 770-772, 775-776 n7; Gray, 2000: 34-68, 144-146; Crowder, 2007: 122, 125-126, 136-137.

Some caution is required here. A weak version of this objection relies on an equivocation about the meaning of 'indifference'. In everyday language, 'indifference' has connotations of equanimity and disinterest. However, the technical use of the term carries no such implications.[23] Being indifferent between S_1 and S_2 simply means believing oneself to have a reason to realise S_1 which is at least as great as the reason one believes one has to realise S_2, and also believing that one has a reason to realise S_2 that is at least as great the reason one believes one has to realise S_1. In Sartre's example, it might be perfectly sensible to say that you are indifferent between your options and yet feel very strongly about them, once we understand that it is the technical sense of 'indifferent' that is intended. Thus, it is not immediately obvious that Sartre's case poses an incompleteness problem for the contractarian model.

There is a stronger version of the objection, however, known as the Small Improvement Argument.[24] Suppose we take Sartre's example again, and say that you are indifferent (in the technical sense) between joining the Free French and looking after your mother – you believe yourself to have as great a reason to do one as the other. Next, suppose that I offer you a cupcake if you will join the Free French. If you like cupcakes, isn't the cupcake an additional reason to join the Free French? And, if the reasons you had were equally balanced before my offer, doesn't the cupcake tip the balance? I take it many people will have the intuition that the prospect of being given a cupcake is not the kind of reason that should be determining what you choose to do in Sartre's case. But this puts an advocate of the completeness condition in a difficult position. Either she has to say that the cupcake does tip the balance, which many people will find counterintuitive. Or, she has to say that it does not tip the balance, and therefore that you may rationally remain indifferent between Sartre's options, even in light of my offer. The problem with the latter position is that it implies a violation of transitivity, and so completeness is retained at the expense of another of the rationality conditions.

[23] Compare Crowder, 2007: 137. See §5.1, *supra*.
[24] Chang, 2002: 667-673; Peterson, 2009: 170-171; Gustafsson and Espinoza, 2010.

The problem is as follows. If we assume that you prefer getting a cupcake and joining the Free French to simply joining the Free French, and that you are indifferent between joining the Free French and looking after your mother, transitivity of preference and indifference requires that you prefer getting a cupcake and joining the Free French to looking after your mother. Saying that you're indifferent between the cupcake-Free French combination and looking after your mother violates transitivity.

Faced with the Small Improvement Argument, we have three options. If we retain the completeness condition, then we either have to, first, bite the bullet in the cupcake case, or second, we have to give up transitivity. Our third option is to give up completeness. If we give up completeness, we can say that, rather than being indifferent between joining the Free French and looking after your mother, the two options are incommensurable.[25] If we don't have to claim that you are indifferent between joining the Free French and looking after your mother, then it no longer follows either that a cupcake should tip the balance between the options or that there is a violation of transitivity. Advocates of the Small Improvement Argument think we should give up completeness. Defending the completeness condition – and hence the contractarian model – requires showing that one of the two other options is more attractive.

Consider, first, the possibility of surrendering transitivity. People with intransitive preferences often strike us as paradigmatically irrational. The classic example illustrating this is the 'money-pump' problem.[26] Suppose Bruce has the following preferences. He prefers eating an apple to eating a pear, eating a pear to eating a peach, and eating a peach to eating an apple:

[25] More precisely, saying that S_1 and S_2 are incommensurable for an agent means that the agent believes the reason she has to realise S_1 is not at least as great as the reason she believes she has to realise S_2, *and* the agent believes that the reason she has to realise S_2 is not at least as great as the reason she believes she has to realise S_1. To use the standard notation: $\neg(S_1\mathbf{R}S_2) \wedge \neg(S_2\mathbf{R}S_1)$ (see n1 and n2).

[26] I owe the following example to Gauthier (1986: 40-41). See also Anand, 1993: 341-343; Cubitt and Sugden, 2001: 123-125; Peterson, 2009: 164-169. For a genealogy of the problem, see Hansson (1993). For a response, see Anand (1987: 200-202).

apple > pear

pear > peach

peach > apple

Bruce's preferences are intransitive – in his case, the fact that S_i (apple) is preferred to S_j (pear), and that S_j (pear) is preferred to S_k (peach), does *not* imply that S_i (apple) is preferred to S_k (peach). Thus, the transitivity condition is violated.

Why does this make Bruce irrational? Suppose Bruce has an apple and some money. Suppose I offer to give him a peach if he will give me $1 plus the apple he has. As set out, Bruce prefers the peach to the apple. Let's assume, for the sake of argument, that the difference between the value he assigns to the peach and the value he assigns to the apple is greater than $1. So, from Bruce's point of view, my offer looks like a good deal, and he rationally makes the trade. He now has slightly less money and a peach.

Knowing that Bruce prefers a pear to a peach, I next offer to give him a pear if he will give me $1 plus his peach (we again assume that, for Bruce, the difference in value between a pear and a peach is greater than $1). As before, the trade makes sense given Bruce's preferences, and so he now has a pear, and still less money. Finally, knowing that Bruce prefers an apple to a pear, I trade him the apple he initially gave me for his pear plus $1. Bruce is now back with the apple he had at the beginning, but is $3 poorer. I can then make the initial offer again, to gain another dollar, and so on until Bruce runs out of money – until I have pumped him dry. "[Clearly] Bruce, who becomes ever less well off as a result of trades, each of which betters his position in terms of his preferences, is engaged in an irrational sequence of choices."[27] Hence, we require that agents have transitive preferences in order to be considered rational.

[27] Gauthier, 1986: 41. Strictly, this example illustrates a problem with intransitivity in the preference relation, not the indifference relation. For the indifference relation version of the argument, to the same conclusion, see Hansson (1993: 478-479).

I discuss objections to the transitivity condition in §5.4. However, for the time being, I take it that the money-pump problem establishes a sufficiently strong presumption in favour of transitivity that we cannot plausibly surrender it. If we can't give up transitivity, our response to the Small Improvement Argument will have to be either to bite the bullet or to give up completeness.

Rather than give up completeness, we should bite the bullet. Here's why: Giving up completeness – allowing the preference sets of rational agents to include incommensurable options – does not necessarily avoid the money-pump problem. Suppose you are on your way to join the Free French. Along the road, I offer you a cupcake. You like cupcakes, so you pay me $1 for the cupcake and continue along your way. A little further on, you reach a fork in the road. If you take the left turn, you will continue on your way to join the Free French, with your cupcake. If you take the right turn, you will go home to look after your mother. However, I have set up a toll booth on the road to the right, and I am charging people one cupcake to pass. So, you now have to choose between joining the Free French and having your cupcake, or looking after your mother. If we assume these options are incommensurable, then it is not irrational for you to choose the road to the right, pay the toll of one cupcake, and return home to your mother. Suppose that is what you do. A little further along the way, you reach another fork in the road. The road to the left will take you to join the Free French, and the road to the right will take you home to your mother (there is no toll booth this time). Since these options are incommensurable, it is not irrational for you to take the road to the left, choosing to join the Free French instead of looking after your mother. Suppose, finally, that a short distance down *that* road you will reach a cupcake stall where I am waiting to deprive you of another dollar. Given incommensurability, this sequence of choices can be repeated until you run out of money, supposedly without you having done anything irrational.

This case is an example of a 'non-forcing' money-pump.[28] There's no rational *requirement* that you make the sequence of choices I describe, so you're not trapped in the way that Bruce is by the 'forcing' money-pump. But each choice in the sequence I describe is, on the incommensurability view, rationally *permitted*. And that looks deeply implausible. Surely (at least) one of the choices has to involve a mistake.

Perhaps it could be objected that the problem is not that there is anything irrational about the agent's incomplete preferences, but instead that the agent is being irresolute.[29] I take it this would be an attempt to block the agent from heading back towards joining the Free French once she has paid the cupcake toll and is on her way home. But look what is now implied. The claim is that, because looking after one's mother has been chosen over joining the Free French and having a cupcake, it cannot later be rational to choose joining the Free French (*sans* cupcake) over looking after one's mother – one has to resolutely stick to one's decision.[30] Why would that be so? Well, one obvious explanation is that, because joining the Free French and getting a cupcake has been chosen over simply joining the Free French, and looking after one's mother has been chosen over the cupcake-Free French combination, it will strike us as weird to then choose joining the Free French (*sans* cupcake) over looking after one's mother. It seems weird because it violates transitivity. The 'resolve' version of the incommensurability view seems to be saying that a rational agent's preferences can be incomplete, so long as she never *acts as if* they're incomplete.

So, in order to avoid the non-forcing money-pump, an advocate of the 'resolve' view has to say that a rational agent's actions must track the actions she would take if she had complete preferences. There are two problems with this.[31] First, on the understanding of rationality I have given, it is a requirement of rationality that (absent

[28] Gustafsson and Espinoza, 2010: 761-762.
[29] McClennen, 1990: 4-17; compare Rabinowicz, 2000: 147-152; Moehler, 2018: 103, 107.
[30] The example assumes there has been no *change* in the agent's preferences. Acting contrary to one's earlier decision is not an example of being irresolute if one has changed one's mind about the decision.
[31] See §5.1.

special circumstances) one's intentional actions be consistent with one's preferences. If an agent is intentionally acting as if she has complete preferences, and if she is rational, then on my understanding she must have complete preferences. If she were rational and had anything other than complete preferences, that would have to be reflected in her behaviour, and on the present view it isn't. Second, on the understanding of preferences I have given, preferences are beliefs about reasons one has for acting. An agent cannot plausibly be rational if she believes she has reason to go along with the non-forcing money-pump. But, if she believes she has reason to avoid the non-forcing money-pump, then implicitly she has complete preferences. Any attempt to evade the non-forcing money-pump thus implies a commitment to completeness. Since giving up completeness leads us into the non-forcing money-pump problem, we should retain completeness.[32] Retaining completeness and transitivity means that, faced with the Small Improvement Argument, we should bite the bullet.

This argument against incommensurability being a plausible source of incomplete preferences will strike some people as specious, particularly when set against the intuition that things like cupcakes simply *cannot* be a deciding factor in cases like

[32] One might worry that the non-forcing money pump proves too much. Suppose one has complete preferences, but that one's preference ordering over certain options includes some indifference relations. Suppose that, when faced with a choice between two such options, one resolves the indifference with a coin toss. This much is standard in decision theory. But note that this procedure might not guarantee that one will avoid going around in circles – if one is (very) unlucky with one's coin tosses, one could get trapped in a loop. One might then think that only strict preference is consistent with rationality, which would be a significant departure from standard decision theory. If we are unwilling to entertain such a departure, we might then be led to reject the argument from the non-forcing money pump. There are three things to say in reply. First, unless there is some cost to getting trapped in a loop, the case given isn't really a case of being money pumped. With no mounting cost, there's no objection to the rational permissibility of proceeding. If getting stuck in a loop *is* costly – in terms of time, say – then it is not clear how the agent will be rationally indifferent in the required way at the relevant choice points (unless she also has intransitive preferences, and is irrational on *that* basis). Second, what is causing the problem in the case described is not the indifference relation *per se*, but the conjunction of the indifference relation and a particular way (repeated coin tossing) of resolving indifference over *sequences* of choices. Even if one thinks the non-forcing money pump implies that some revision of standard decision theory is necessary, such a revision need not take the form of excluding rational indifference (indeed, it certainly should not). Finally, if pressed, I would still not be inclined to abandon the non-forcing money pump at this juncture. If standard decision theory says it is rational to pursue costly loops which are equivalent to a non-forcing money pump, that seems to me to be a good reason to revise (or extend) standard decision theory. Nonetheless, any such revision should address procedures for resolving indifference across sequences of choices, rather than reject rational indifference.

Sartre's. In order to bolster the argument, we need some explanation of this intuition which does not lead us to incommensurability. But I think such an explanation can be given.

Suppose I offer you a choice between two boxes, *A* and *B*. One of the boxes contains $10,000; the other is empty. However, you don't know which is which. You have to choose a box. If you choose the box with the money, you get to keep it. Given the circumstances, I take it you have as much reason to choose one box as the other. Suppose I then say that, if you pick *A*, I will add a $1 bonus to whatever is in the box. Does this give you a reason to prefer *A* to *B*?

On a standard account of choice under uncertainty, and lacking any further information, we would assume that each box is equally likely to contain the money. Including the bonus I've offered, the expected value of choosing *A* is thus $5,001. The expected value of choosing *B* is $5,000. Therefore, it is rational to choose *A*. Note, however, that what the standard account is doing is telling us how to choose rationally in conditions of uncertainty. It is not telling us which box is more valuable. Suppose it is actually *B* that contains the money. Objectively, then, you have most reason to choose *B*, not *A*, regardless of my offer. The bonus I have offered isn't large enough to be relevant to the question of whether *A* is better than *B*. Given what you know about the boxes, the bonus is not enough to compensate for having chosen the wrong box. Objectively, the small improvement is irrelevant, and you know that it's irrelevant, regardless of where the money is.

Similarly, in cases like Sartre's, I take it the problem is not that our options are incommensurable, but that we are making decisions under a significant degree of uncertainty.[33] When faced with the Small Improvement Argument, our intuitive rejection of the claim that small improvements resolve indifferences is driven by the

[33] Broome (1978: 330) resists this conclusion by stipulating that the agent is fully informed about the likely consequences, having had long experience with each of the two kinds of life between which she is choosing. In such a case, however, I find myself lacking any intuition whatsoever in favour of incommensurability. And even if, given the stipulation, one were still to think that the options are incommensurable, the non-forcing money-pump seems to me to offer an entirely satisfactory reason to reject one's intuition.

belief that (i) there is a fact of the matter as to which of our options is better, and (ii) the gap between the value of one option and the other is not plausibly so small that the small improvement will actually bridge it, and so (iii) it seems to make no sense to let the small improvement decide what we believe we ought to do. Thus, we can explain our intuitive response to the Small Improvement Argument without there being any reason to grant incommensurability, because we plausibly have the same intuition in the box case, without the box case involving any objects of incommensurable value.[34]

It may be thought that there is a disanalogy between the box case, which involves prudential considerations, and Sartre's case, which involves moral considerations. However, we can make the two more similar by setting up the box case so that the money you win will be given to a charity doing important work, and so you have a moral reason to try to win as much money as possible.

Alternatively, perhaps the problem with the cupcake offer is not that it is a *small* improvement, but rather that it is the wrong *kind* of improvement – perhaps the thought is that prudential improvements should not weigh at all in moral decisions of the kind Sartre describes. However, it is trivial to reformulate the case in such a way that what is at issue is a small moral improvement. Furthermore, to support the incommensurability view, the claim would have to be that prudential considerations (such as receiving a cupcake) and moral considerations (fidelity to one's family or one's country) are incommensurable. This must not be confused with the claim that moral considerations always trump prudential considerations. This second claim actually relies on completeness – it claims that moral and prudential considerations *are* commensurable, and that the former always outweigh the latter. Whatever we think of that view (and I have argued against it), it is much more radical to say, when asking if I have greater reason to join the Free French or to get a cupcake, that there is no fact of the matter. But that is what is required by the claim that moral and

[34] Note that this argument doesn't actually depend on it being *rational* to disregard the bonus offer in the box case. My purpose is simply to explain the intuition we have in the cupcake case without having to appeal to incommensurability. That intuition need not be rational for the argument to go through – people have irrational intuitions all the time.

prudential reasons are incommensurable. It is difficult to definitively rule out the possibility that prudential and moral reasons are incommensurable, but I proceed on the basis that they are not.[35]

Thus, neither the cognitive limitations of real agents nor the suggestion of incommensurable options gives us compelling reasons to reject the completeness condition upon which the contractarian model depends. Given the prescriptive function of the model, idealising the cognitive capacity of the agents is not problematic. And, while the Small Improvement Argument shows that completeness can lead to some counterintuitive implications, the alternative – incommensurability – has implications that are even less attractive.

§5.4 Objections to transitivity

Just as it was said that real people don't have complete preferences, it is undeniably the case that real people sometimes exhibit preferences that appear to be intransitive.[36] In itself, however, this isn't sufficient to ground an objection to the transitivity condition. The contractarian model purports to capture the behaviour of rational agents, and the fact that real people sometimes behave irrationally is no reason to think that they (prudentially-)ought to behave in that fashion. In order for an objection to have some bite we need a plausible example of an agent with intransitive preferences whom we nonetheless think is rational. If such an agent exists, then the transitivity condition is an inappropriate way of modelling rational agents.

The money-pump argument described in §5.3 gives us reason to doubt that agents with intransitive preferences are rational. Nonetheless, it has been argued that they can be. The most persuasive objections to transitivity typically come in one of two forms (or in a form that combines aspects of each). There are *discernment* objections

[35] See §2.4.
[36] See Anand (1993: 337) for a survey of empirical literature examining this.

and there are *multidimensionality* objections.[37]

§5.4.1 Discernment objections

Consider the following example of a discernment objection.[38] Suppose I can choose between a thousand cups of tea. One cup, C_0, has no sugar in it. A second cup, C_1, contains a single grain of sugar. The next cup, C_2, contains two grains of sugar, and so on up to C_{999}. For any C_n, I am plausibly indifferent between C_n and C_{n+1}, because my faculty of taste cannot discriminate between cups which differ by a single grain of sugar. Thus, I am indifferent between C_0 and C_1, between C_1 and C_2, and so on. However, I am not indifferent between C_0 and C_{999}. I find the latter too sweet. This is a violation of transitivity in the indifference relation, but it seems odd to conclude that I am irrational.

The natural response to this objection is to say that there is some C_n which is a tipping point – at some n grains of sugar, the tea becomes too sweet for my taste. I am indifferent between any pair of C_0, C_1, C_2... and C_{n-1}, but I am not indifferent between C_n and any of the cups up to C_{n-1}. If this were so, transitivity would be preserved. This response, however, relies on me being able to tell the difference between levels of sugar which differ by a single grain; but presented with C_{n-1} and C_n, isn't it implausible to think that I will be able to distinguish them, and hence that I will have any (strict) preference between them?

Suppose we approach the problem in a slightly different way.[39] Suppose you pick a

[37] For discernment objections to transitivity, see Schumm (1987), Quinn (1990), Aldred (2007), or Hansson and Grüne-Yanoff (2017: §2.3). For multidimensionality objections to transitivity see Schumm (*op. cit.*), Bar-Hillel and Margalit (1988), Philips (1989), Anand (1993: 344), Temkin (1996), Rachels (1998), Aldred (*op. cit.*), or Hansson and Grüne-Yanoff (*op. cit.*). Anand (1993) provides a helpful survey and evaluation of objections to the transitivity condition, including a number I do not have space to consider here. I should note, in particular, the form of objection pressed by Rachels and Temkin, which has attracted attention in mainstream moral and political philosophy. While provocative, I believe it can be answered effectively – see Binmore and Voorhoeve (2003).
[38] Hansson and Grüne-Yanoff, 2017: §2.3.
[39] This solution is a version of the triangulation method suggested by Quinn (1990: 81).

cup, C_r, from the set at random and give it to me to taste; you then ask me if I find the tea too sweet, and note my reply. Suppose you repeat this procedure with a range of cups. Some of them I find too sweet, some of them I don't. Over repeated iterations of the procedure, we will be able to partition the set of cups into those I find too sweet and those I do not. Assuming my tastes are consistent (such that, presented with some C_r, I always place it into the same category, either too sweet or not too sweet), and that at least one cup is too sweet, this implies that there will be a sharp line drawn at some point in the set, and therefore that I *can* distinguish between cups which differ in sweetness by just a single grain of sugar. What is going on here?

The thought is something like the following. It may be that I am incapable of discriminating between some adjacent cups, C_{n-1} and C_n. Nonetheless, I can compare each of the two options to an index case – C_0, say. Cups close to C_0 will not be distinct from it, as far as I can tell, but I can certainly tell the difference between C_0 and a cup that is sufficiently far from C_0. As I move further from C_0, there will be some cup, C_n, at which I first become capable of distinguishing the test cup from the index cup. So, even if I can't directly distinguish between C_{n-1} and C_n, I can tell that one of them is distinct from C_0, to my taste, and the other is not. In this way, I can precisely partition the set of cups. With the partition in place, I no longer have intransitive preferences. As before, I am indifferent between C_0 and all of the cups up to C_{n-1}; I am indifferent between C_n and all of the remaining cups up to C_{999}; and I prefer C_0 to C_{999}. But because I prefer C_{n-1} to C_n, there is no intransitivity. An indifference relation holds within each partition, and a preference relation holds between cups from different partitions.[40] The discernment objection relied on me being indifferent between C_{n-1} and C_n, but by triangulation with C_0, I can say that a preference relation holds between them instead, dissolving the objection.

Warren S. Quinn considers a solution of this kind to the discernment problem and rejects it. The case he is dealing with is a little different.[41] He asks us to imagine that

[40] Quinn seems to have something like this method in mind when he discusses what he calls "filtered series" (*ibid*.: 86).
[41] 1990. I have made minor adjustments to the case, for ease of exposition. I also focus here on the discernment aspect of Quinn's example, but it can be understood as a case of multidimensionality as

there is a certain device which can cause you to experience pain. The device has 1,001 settings, ranging from S_0 (when the device is switched off), to S_{1000} (the maximum setting). Once a week, you have a choice. You can either leave the device at its present setting (S_n), or you can increase the setting by one step (move to S_{n+1}). Every time you increase the setting, you receive $10,000. However, you may never reduce the setting. Importantly, the device is tuned finely enough that you cannot detect a difference in your degree of discomfort between any two adjacent settings. Nevertheless, you are perfectly capable of detecting differences between settings that are sufficiently far apart, and there are some settings (including S_{1000}) which cause you so much pain that you would forego all of the money accumulated in reaching those settings in order to return the device to S_0.

As before, Quinn's 'self-torturer' case sets up a problem in which an agent has intransitive preferences. She prefers any S_{n+1} to any S_n, but also prefers S_0 to S_{1000}:

$$S_0 < S_1 < S_2 \ldots < S_{999} < S_{1000} < S_0$$

Moreover, Quinn claims – quite plausibly – that the self-torturer is rational. Given a device that can be tuned very finely, it is plausibly the case that the self-torturer will be incapable of distinguishing between any two adjacent settings. If that is so, then each week she will rationally prefer to be given $10,000 in exchange for no discernible cost in terms of her level of discomfort. This will, of course, lead her to S_{1000}, and it is again plausible to suppose that she would prefer to be without any of the money, but back at S_0, rather than being at S_{1000}.

Why can't the triangulation method solve this problem, as it did earlier? Quinn tells us:

> "[Surely] it ought to be an open empirical question whether such triangulations are possible. If there are increments of voltage just small

well, with money as the second dimension. See §5.4.2, *infra*.

> enough to be directly undetectable, it seems there might be even smaller increments that cannot be detected by triangulation. And I want such a case. The self-torturer (and his observers) try to triangulate differences between adjacent settings but honestly cannot – not because they are inattentive, but because we have made the increments of current too small to make *any* difference in comfort, even one that can be detected only by triangulation."[42]

This response to triangulation looks like a mistake, however. It proves too much – it turns a sorites paradox into one of Zeno's paradoxes.[43] In order for Quinn's objection to have force, we must prefer S_0 to S_{1000}. Otherwise, we cannot generate the intransitivity. So, it cannot be the case that we constantly increase the setting on the device, but in such a way that the *cumulative* increase remains undetectable relative to S_0. At some point, we have to actually *arrive* at a state in which we are in pain. Which is to say, it must be the case that there is at least one S_n that is sufficiently far from S_0 that we can distinguish the two. If this were not so, there would be no problem, since it would mean we wouldn't be able to distinguish between S_0 and S_{1000}, and could happily pocket all of the cash and go about our business. For any particular S_n, then, it will either be the case that we can distinguish it from S_0 or that we cannot distinguish it from S_0, and this is true no matter how finely grained the increments between the settings are. This being so, we can partition the set of options as before, and there will be no intransitivity – at the boundary between the partitions, the preference relation will reverse.

§5.4.2 Multidimensionality objections

The second prominent set of objections to transitivity relies on cases in which states of affairs must be evaluated on more than one dimension. The classic example of a multidimensionality objection is attributed to G. H. von Wright. I take this version of it from Jonathan Aldred:

---

[42] 1990: 81, emphasis in original.
[43] Compare Binmore and Voorhoeve, 2003: 276-277.

> "At a dinner party you are to be offered some fruit. If offered an orange and a small apple [from which to choose], you will pick the orange. If offered an orange and a large apple, you will pick the apple. But if offered a large apple and a small apple, politeness demands that you take the small apple. Transitivity would have you pick the large one."[44]

As the problem is constructed, you appear to have the following intransitive preferences:

large apple > orange > small apple > large apple

However, as Aldred notes, this kind of problem is no problem at all for transitivity. The introduction of politeness means there is a second dimension of value over which you have preferences. Once we make this explicit, your composited preferences over the two dimensions are transitive:

large apple, be polite > orange > small apple > large apple, fail to be polite

Consider a more complicated version of this kind of objection.[45] Smith is shopping for Christmas decorations and is choosing between three available boxes. Each box contains a red ball, a blue ball, and a green ball. In Box *i*, we designate the red ball R_i, the blue ball B_i, and the green ball G_i. Suppose Smith has the following preferences:

$R_1 > R_2 > R_3$
$B_2 > B_3 > B_1$
$G_3 > G_1 > G_2$

Thus, his favourite red ball is in Box 1, his favourite blue ball is in Box 2, and so on.

[44] 2007: 379.
[45] The following example is adapted from one given by Schumm (1987).

When we now compare the boxes, rather than the balls, intransitivity threatens. Suppose Smith is comparing Box 1 and Box 2. He prefers Box 1 on the red and green dimensions, but prefers Box 2 on the blue dimension. Suppose he decides that Box 1 is preferable, all-things-considered, on the basis that it is preferable on more dimensions than Box 2. Suppose, next, that Smith is comparing Box 1 and Box 3. He prefers Box 1 on the red dimension, but prefers Box 3 on the blue and green dimensions. Applying the same reasoning as before, he decides that Box 3 is all-things-considered preferable to Box 1. Finally, suppose that he is comparing Box 3 and Box 2. He prefers Box 2 on the red and blue dimensions, and prefers Box 3 only on the green dimension. Thus, as before, all-things-considered he prefers Box 2 to Box 3. Now, however, his all-things-considered preferences over the boxes are intransitive:

Box 1 > Box 2 > Box 3 > Box 1

This shouldn't be surprising, since the example is simply a redescription of a classic Condorcet voting cycle, with each box serving as a candidate, each dimension serving as a voter, and with a majority decision rule.[46] Clearly, given his preferences, Smith can be money-pumped in the standard manner. Moreover, unlike von Wright's fruit example, in Smith's case making the various dimensions explicit does not immediately resolve the intransitivity.

In order to defend transitivity as a requirement of rationality, we must be able to point to something irrational about Smith. He can hardly be thought to be irrational in his basic (transitive) preferences over the balls. The problem is the manner in which these dimension-by-dimension preferences are integrated into an all-things-considered judgement. The majority decision rule – the principle that, if one prefers *i* to *j* on more dimensions than one prefers *j* to *i*, one therefore prefers *i* to *j* all-things-considered – can result in cyclical preferences. If we once again insist that rational

[46] Bar-Hillel and Margalit, 1988: 125-127; Rabinowicz, 2000: 126-129.

agents must not be vulnerable to a money-pump, we are led to the conclusion that rationality constrains the manner in which an agent may integrate dimension-by-dimension preferences. Thus, Smith is irrational because he is using the wrong kind of integration rule to assemble his all-things-considered judgements.

An alternative integration rule would have allowed Smith to avoid the money-pump. Suppose, rather than simply ranking his preferences over the various balls ordinally, Smith were to assign a cardinal value to each ball:

R_1 (\$3) > R_2 (\$2) > R_3 (\$1)
B_2 (\$4) > B_3 (\$3) > B_1 (\$2)
G_3 (\$5) > G_1 (\$3) > G_2 (\$1)

Thus, R_1 is worth \$3 to him; B_1 is worth \$2, and so on. We can now express Smith's preferences over the boxes using the sum of the values he ascribes to the individual components of each box:

Box 1 (R_1, B_1, G_1) = \$8
Box 2 (R_2, B_2, G_2) = \$7
Box 3 (R_3, B_3, G_3) = \$9

Smith's preferences are now straightforwardly transitive:

Box 3 > Box 1 > Box 2

Because it maps each combination of components onto a scale with a single dimension, this approach to the integration of dimension-by-dimension preferences guarantees that Smith's all-things-considered preferences over the boxes will be transitive.[47] Each box will be assigned a number on a scale of value (in this case represented by dollars), and Smith will prefer one box to another, or will be

[47] Bar-Hillel and Margalit, 1988: 124-125.

indifferent between two boxes, according to the value each is assigned. Call this an *independent pricing procedure*.[48] In order to be sustained, a multidimensionality objection to transitivity must deny either the possibility or the validity of implementing an independent pricing procedure in some particular case.

There are two kinds of cases in which we might think that an agent could fail to have the cardinal preferences upon which an independent pricing procedure relies, and might nonetheless be rational. First, it might be the case that the dimensions which must be assessed are incommensurable. This would make the independent pricing procedure impossible. However, as I have argued above, if Smith has incommensurable preferences, he will be vulnerable to a non-forcing money-pump, and thus we can conclude that he is to that extent irrational.[49] Second, it might be the case that the dimensions of value are not independent of one another – if, for example, the value of a particular shade of red depends on the shade of blue with which it is combined. This would violate the monotonicity requirement.[50]

It is clearly the case that some dimensions of value are interdependent – the case of matching shades of colour is a plausible example. Such a possibility, however, does not invalidate or prevent the application of an independent pricing procedure. The procedure requires that, when aggregating multiple dimensions of value, an independent price be assigned to each independent dimension. In Smith's case, if the red and blue dimensions are interdependent, but the green dimension is independent of the red and blue dimensions, the procedure will assign one price to the red-blue composite dimension and another price to the green dimension. If all three dimensions of colour are interdependent, Smith's case will not be a case of multiple dimensions – the procedure will assign a single price to combinations of red, blue, and green. Thus, the possibility of interdependence presents no special problem for the application of the independent pricing procedure either.

[48] *Ibid*.: 124.
[49] §5.3.
[50] §5.2. For some further remarks on the relationship between the transitivity and monotonicity requirements see Binmore and Voorhoeve (2003: 279).

I take it incommensurability and interdependence exhaust the plausible ways in which an agent might initially appear to be rational but nonetheless be unable to apply, or unjustified in applying, an independent pricing procedure. In the case of incommensurability, I've argued that the agent is not actually rational. In the case of interdependence, I've argued that an independent pricing procedure can validly be applied.

§5.4.3 Money-pump redux

Neither discernment nor multidimensionality objections give us reason to reject transitivity as a condition of rationality. Discernment objections can be resolved using triangulation, and Quinn's reasons for rejecting triangulation are unpersuasive. Moreover, discernment problems look like artefacts of constructing examples involving the physical senses; it's not obvious that any similar problem attends preferences which are purely intellectual – constituted by beliefs about reasons for action. Multidimensionality objections can be resolved using an independent pricing procedure. The only instance in which an independent pricing procedure would not be viable would be in the presence of incommensurable options. But I have argued that incommensurability is inconsistent with rationality.

By way of concluding this discussion of the formal conditions of rationality, let me anticipate a possible worry about the manner in which the arguments of §5.3 and §5.4 have proceeded. In §5.3, I argued that our commitment to transitivity means we cannot give up completeness in the face of the Small Improvement Argument. In §5.4.2, I have relied on the rejection of incommensurability to defend transitivity. The rejection of incommensurability, in turn, relied on the success of the argument for completeness. Since the argument for completeness has apparently relied on a commitment to transitivity, and the argument for transitivity has apparently relied on a commitment to completeness, it may appear that, taken together, the argument of the two sections is circular.

There are two responses to this worry. First, note that the circularity is only apparent. In both sections of the argument, my ultimate appeal has been to the money-pump problem. I reject incommensurability because it leads to the non-forcing money-pump, and rejecting incommensurability leads me to accept completeness. The money-pump problem is the positive argument I offer in favour of transitivity – it does not itself rely on the acceptance of transitivity. Thus, the arguments for both completeness and transitivity rely on the money-pump, and not on one another.

The second response is to say that even if my argument were circular, the circularity would not be vicious. In reflective equilibrium, it speaks in favour of completeness and transitivity (and monotonicity and continuity) that they cohere with one another. That they stand and fall together is a reason to accept each, since it means that the cost of rejecting one is all the higher.

It is worth noting that transitivity is not the only possible condition of rationality which could be derived from the money-pump. Transitivity is sufficient to avoid the money-pump problem, but it is not necessary.[51] Thus, we could accept completeness but pair it with something other than transitivity. Nonetheless, I have also argued that the most common other suggestion – something like a resolve condition – is not a viable alternative, given the understanding of preferences and intentional action which I have described.[52] As I said at the outset, a full discussion of the formal conditions is beyond my scope. But it seems to me that they can be defended against the most persuasive objections to which they are usually subject. I will proceed, therefore, to take them for granted.

§5.5 The circumstances of contractarian justice[53]

The third element of the contractarian model consists of a set of facts about the world

[51] Cubitt and Sugden, 2001.
[52] §5.1, §5.3.
[53] An earlier version of this section formed part of my MPhil thesis (2006).

in which the model's agents are acting. I will discuss five of them here. I will first state and explain the five propositions, before arguing that each is a reasonable assumption to make in building the model.

§5.5.1 Five facts about the world

First, the agents are not alone. The world contains more than one agent, and the agents know this.

Second, the agents find themselves in circumstances of 'moderate scarcity'.[54] This means that it is not possible simultaneously to fully realise all of the goals held by all of the agents. Relative to what the agents (in the aggregate) want, resources are *scarce* (though they might not be scarce relative to what any individual agent wants). However, although resources are scarce, the scarcity is *moderate*, such that the realisation of some part of each agent's goals is consistent with the realisation of some part of the goals of each other agent. So, for example, resources are not so scarce that the agents are fighting over a single lifejacket on a sinking ship.

Moderate scarcity is a weak condition. It excludes only circumstances of absolute abundance, and circumstances in which there are agents who have *no* goals which can simultaneously be realised. A broad range of possible circumstances lies between those extremes. Note that the moderate scarcity condition has both an 'objective' (agent-independent) component and a 'subjective' (agent-dependent) component.[55] Scarcity is defined in terms of the resources actually available in the world, but also in terms of what the agents (in the aggregate) want to do with those resources. Thus, if we were playing around with a model, we could move from a case of moderate scarcity to a case of abundance by two means: either by increasing the resources available to realise goals, or by altering the goals so that they can all simultaneously be realised (and, similarly, we have two means of moving from

[54] Rawls, 1999a: 110.
[55] *Ibid*.: 109-110; Gauthier, 1986: 113-114; Vanderschraaf, 2019: 96.

moderate to absolute scarcity). Note also, then, that 'abundance' strictly includes states of affairs in which there is such harmony in agents' ends that, even with very few resources available, everyone agrees on what to do with them. Moderate scarcity rules out this possibility – by assuming moderate scarcity, we assume that there are competing and inconsistent claims on at least some resources.[56] A 'resource' is here understood as anything in the world that is the object of an agent's preferences (including other agents, insofar as they or their actions figure in the preferences of some agent). Finally, note that adding agents to the world, or eliminating them from it, would have an effect on the model equivalent to altering the resources or the goals in the world. Agents in the model are both sites of ends and means of realising ends – adding and subtracting agents can thus affect both the subjective and the objective components of the moderate scarcity condition.

Third, the agents are capable of intentionally interacting in a manner which affects the realisation of one another's goals. Imagine you and I have been shipwrecked, but in escaping the sinking ship we have ended up on different desert islands. We are aware of one another's existence, but cannot interact in any way. Each island has sufficient resources for us to survive, but neither has sufficient resources to completely realise our goals. This world meets the first two conditions, but it is not the kind of world I am modelling. In the contractarian model, agents are capable of interacting with one another, and the degree to which the goals of each agent are realised depends, to some extent, on how each other agent decides to act.[57]

So, there are agents, they can interact with one another in such a way as to affect the realisation of one another's goals, and the world is not so harsh that coordination of their actions is pointless, nor so idyllic that it is unnecessary. These three conditions together open up the possibility of conflict, as the agents compete for scarce resources. But they also create a space within which the agents can coordinate to their mutual advantage, because if the agents can coordinate their actions in particular ways, the goals of each agent can be realised to a greater degree

[56] Rawls, 1999a: 110. See also Leys, 1965: 274; Binmore, 1994: 55-56; Margalit, 2009: 194.
[57] Rawls, 1999a: 109-110; Gauthier, 1986: 114; Vanderschraaf, 2019: 3, 43, 114.

than they would be in the absence of coordination. Circumstances in which these three conditions hold are typically termed the 'circumstances of justice'; not because they are just, but because they are the circumstances in which it is plausibly thought that principles of justice can be relevant to the question of what we (morally-)ought to do.[58]

The conclusion of the contractarian argument is supposed to be a prudential prescription to act in whatever is the most effective way of realising one's goals, given the circumstances in which one is acting. Our question now is, given that the agents in the contractarian model are rational, and find themselves in the circumstances described, what is the most effective means by which they can realise their goals?

One possibility is violent conflict. But there are many costs to engaging in conflict. These include the risk of injury or death, the risk of misjudging the outcome of the contest, the alienation or destruction of a potential ally, the opportunity cost of spending strength against one rival as opposed to others, and the possibility that engaging in violent conflict is inconsistent with one's moral commitments.[59] Further, there is no guarantee that a conflict between agents will be decisive – rather, it might drag on in some kind of stalemate. If there were means of achieving one's ends without being exposed to these costs and risks, it seems plausible to say that, for most agents, those means would be preferable to engaging in conflict.

The alternative to violent conflict is a negotiated agreement on some coordinated plan of action. There are costs here, too. First, there are transaction costs associated with coming to an agreement. If there are many agents, with complex goals, these could be substantial. Moreover, some ends are time-sensitive, such that if one does not act quickly one will not be able to realise them, and securing agreement takes time. However, the greatest cost of agreement as an alternative to conflict is likely

[58] See §1.5 n37, and §1.7; Rawls, 1999a: 109-112, and especially 109 n3; Moehler, 2018: 97; Vanderschraaf, 2019: 85-99. Tellingly, these are also the circumstances which define a bargaining situation – see §6.1, *infra*; compare Vanderschraaf, 2019: xiii, 110, 112, 275.
[59] Abizadeh, 2011: 313.

to be that one (probably) won't get everything one wants. Take agents *L* and *M*. In order for *L* to secure *M*'s agreement to a coordinated plan, *L* will (probably) have to offer *M* some incentive to coordinate his actions with *L*'s in such a way as to realise *L*'s goals. Providing this incentive is a cost to *L*. (Moderate scarcity rules out the fortuitous coincidence of *L* and *M*'s goals in general, if not in any particular case, and so it rules out the possibility that *M* needs no incentive to coordinate, or that providing the incentive is not a cost for *L*.)

The question, then, is which route carries the lower costs, and the greater chance of success. The fourth fact about the world which the contractarian model assumes is that agreement is the more effective of the two available means. This assumption, of course, still leaves the content of the agreement open.

The fifth fact about the world which the model assumes is a 'statist' claim. This is the claim that spontaneous order – anarchy – will not be the most effective means of realising the goals of the agents in the model. Rather, their goals (and perhaps their nature) will be such that the goals will most effectively be advanced by some form of political institutions. For example, absent very optimistic assumptions about human behaviour, a preference for states of the world in which one is not attacked by other agents over states of the world in which one is attacked, will give most agents a reason to agree to support at least a minimal set of institutional arrangements.[60] While this claim strictly rules out the *absence* of political institutions, I intend it to leave the *form* of political institutions otherwise open.

§5.5.2 Evaluating the facts

These are the five facts about the world the model will take for granted: knowledge that there are multiple agents in the world, moderate scarcity, the capacity for interaction, the relative effectiveness of agreement, and the relative ineffectiveness

[60] Wellman, 1996: 229-232; compare Locke, *The Second Treatise of Government*, Chapter 9 (2005: 71-74).

of anarchy. Are they reasonable assumptions to make? In order to be reasonable assumptions, they don't have to perfectly reflect the real world. Given that the contractarian model is prescriptive, some idealisations are necessary. Other departures from an accurate reflection of the world might be justified as worthwhile simplifications, as reasonable presumptions in conditions where we lack full information, or as irrelevant to the features of the world the model is designed to capture.[61] I take it the first of the assumptions – knowledge that there are multiple agents in the world – is uncontroversial. Each of the other conditions, however, deserves some further consideration.

On its face, moderate scarcity does seem like a weak condition, and a reasonably accurate reflection of the world in which we find ourselves. Our world is not a sinking ship; society is not a zero-sum game. You and I can coordinate in ways that leave us both better off. But we also don't live in conditions of abundance. We aren't indifferent about how the fruits of our coordination are divided.[62]

The condition is somewhat stronger than it looks, however. It claims that productive coordination is possible between *any* agents in the world. And that might not be true. Suppose *L* hates *M* – she would like nothing more than to kill him, and if she cannot accomplish that, wants him to suffer as much as possible, and to frustrate all of his plans. Suppose these preferences dominate any other preferences she has (including even preferences about her own preservation). We can call preferences of this kind *pathological* preferences.[63] Although agents with such preferences are rare, they do exist. In reality, many of them probably fail to meet one or more of the rationality conditions, but there is no formal inconsistency between being rational and having the destruction of other agents as one's highest end. Regardless of how *M* feels about *L*, if *L* ranks *M*'s destruction sufficiently high in her preference set, productive coordination between them will be impossible – *L*'s preferences are

[61] See the discussion in the Interface.
[62] Rawls, 1999a: 4, 109.
[63] I take this term from my MPhil thesis (2006). Compare Hobbes's discussion of the fifth law of nature, in Chapter 15 of *Leviathan* (1996: 106).

defined in such a way that gains for *M* are necessarily losses for *L*.[64]

It's worth emphasising just how extreme *L*'s preferences have to be in order for her to fall outside the bounds of moderate scarcity. If she were to prefer her own survival over killing *M*, productive coordination between them would be possible – they could both, in principle, gain from an agreement not to try to kill one another. To be clear, the problem posed by agents with pathological preferences is not that they want to kill other agents, or that their end is to frustrate the realisation of others' goals, but that they have these ends *and* prefer them over all others. We can coordinate productively with Nazis, so long as they're Nazis who have a healthy regard for their own lives.

If there are agents with such extreme preferences in the world, then in assuming moderate scarcity, the contractarian model is to that extent inaccurate. Nonetheless, we shouldn't revise the moderate scarcity condition at this stage. The existence of agents with pathological preferences implies that not all agents can be included in a coordination agreement – in our example, we can include either *L* or *M*, but not both. So, one question we will want to answer is which of *L* and *M* gets included. But that doesn't mean we won't be interested in the question with which we started. *Given* some included population of rational agents, among whom productive coordination is possible, what is the most effective means by which they can realise their goals?[65] We would only have reason to dispense with that question, and hence with the moderate scarcity condition, if there were *no* agents among whom productive coordination were possible. But that's radically implausible. So, the possibility of agents with pathological preferences is not a reason to reject the moderate scarcity condition, it simply draws our attention to the fact that the model will need some principles governing the inclusion of agents in cases where not all agents can be included.[66]

[64] Note, however, that the concept of a pathological preference is formal, and so relative. That is to say, from the point of view of *L*, it's *M* who has pathological preferences, inasmuch as his preference for survival over death is strictly incompatible with her preference for his death over his survival.
[65] See the discussion of the contractarian constituency of justification in §3.2.1.
[66] The most obvious solution is that we (prudentially-)ought to include whichever of *L* and *M* is more productive – whichever has the capacity to better contribute to the realisation of other agents' goals.

A similar kind of issue arises with respect to the interaction condition. The interaction condition supposes that each agent is capable of affecting the degree to which the goals of each other agent are realised. But that might not be so. David Hume provides the classic example:

> "Were there a species of creatures, intermingled with men, which, though rational, were possessed of such inferior strength, both of body and mind, that they were incapable of all resistance, and could never, upon the highest provocation, make us feel the effects of their resentment; the necessary consequence, I think, is, that we should be bound, by the laws of humanity, to give gentle usage to these creatures, but should not, properly speaking, lie under any restraint of justice with regard to them..."[67]

We need not go so far as to imagine a whole species. The interaction condition is violated if there exist any two agents in the world who differ sufficiently in the extent of their capacities that one cannot affect in any way the degree to which any of the other's goals are realised. Note that, as in the case of moderate scarcity, the interaction condition has two aspects. It could be that *L* cannot act in such a way as to affect the realisation of any of *M*'s goals because *L* is, like Hume's creatures, very weak relative to *M*. But, it could also be that *L* cannot affect the realisation of any of *M*'s goals because *M* has very unusual goals. Suppose, for example, that *M* has *no* goals – perhaps he is a (caricature of a) Stoic philosopher or a Buddhist monk and has rid himself of all desires. In such circumstances, *L* could have an enormous capacity to affect the world, without being in a position to have the slightest effect on any of *M*'s goals.

As with the moderate scarcity condition, the possible existence of agents who don't

---

In the event that this is *L*, obvious moral objections follow (see Moehler, 2018: 86-87). I will discuss this problem in Chapter 8. The problem doesn't ground any objection to the contractarian theory at *this* stage, however, since my concern right now is whether the five assumptions are reasonable for the purpose of building the model, not whether the model's prescriptions are defensible.

[67] *An Enquiry Concerning the Principles of Morals*, §3.1.18 (1998: 88). Compare Barry, 1995: 42.

meet the interaction condition is not a reason to revise the condition. Clearly, we live in a world where lots of people can affect the realisation of at least one of our goals – anyone who can trip me, or throw a pebble at me, or make an annoying sound can affect the realisation of my goals. Meeting the interaction condition doesn't have to involve affecting the realisation of my goals to a large degree.[68] Agents with very odd preference sets, such as *M*, may exist; agents with very limited capacities, such as *L*, do exist. But the existence of agents of these types does not undermine a model based on assuming their absence, because their presence doesn't change the coordination problem that the agents who do meet the interaction condition have to solve.[69]

The assumption that agreement will be more effective than conflict at advancing the agents' goals initially appears to be the most vulnerable of the claims I have made for the contractarian model. Why would it be reasonable to assume that negotiated agreement carries lower net costs or has a greater prospect of success than conflict? First, it is important to be clear what this assumption is claiming. 'Agreement' does not mean 'uncoerced agreement'. Handing over your money because the highwayman has threatened your life is a form of agreement. Agreement is here contrasted with brute force. And so the model's claim is that agreement, including agreement motivated by threats, is likely to be more effective than recourse to brute force.[70]

With this clarification in place, consider the costs on either side of the question. The costs of conflict, I have said, are the risk of injury or death, the risk of misjudging the outcome of the contest, the alienation or destruction of a potential ally, the opportunity cost of spending strength against one rival as opposed to others, the possibility that engaging in violent conflict is inconsistent with one's moral commitments, and the risk that the use of force will not be decisive. The costs of agreement are transaction costs, time-sensitivity costs, and the cost of conceding to

[68] Binmore, 1994: 74.
[69] I return to the problem of agents who fail to meet the interaction condition in §8.5.
[70] Compare §2.1.

others some share of the proceeds of coordination.

For the purposes of the model, I propose to neglect transaction costs.[71] While these are significant in the real world, the point of the model is to tell us what the content of the agreement between the agents (prudentially-)ought to be. In the same way that the poker model tells me what a player without my cognitive limitations would do in my circumstances, the contractarian model is supposed to tell agents what they would agree to if they were not limited by, for example, the time it takes to negotiate complex agreements.

Beyond that, it is going to be an empirical question as to whether conflict or agreement is a better strategy for any particular agent. But there are good reasons to doubt that agents with typical preferences would be better off pursuing a conflict strategy. Consider the kind of agent for whom a conflict strategy would be most plausible – an agent with a high capacity for violent conflict. It is possible that preemptive attacks on some other agents would yield the aggressor some early gains, but the medium-term prospects for such an approach seem dim. How long would other agents sit around waiting to be picked off one by one? And, even if they did, depopulating the world of agents reduces the resources in the world available to realise one's goals – from a brutally amoral point of view, slaves are more useful than corpses, but slavery arises from agreement, not brute force. There is thus a limit to how effective brute force can plausibly be. Finally, note that agreement and conflict are not strictly alternatives. Pursuing an agreement strategy does not foreclose the use of force later – the whole point of making threats is to indicate the future circumstances under which one intends to use force. While initial conflict doesn't strictly rule out subsequent agreement (and, of course, the communicative effects of the use of force should not be underestimated), it does seem likely that the use of force will make it more difficult subsequently to establish the trust upon which any form of agreement will rely.[72] Absent special circumstances, threats are a shrewder approach than brute force. They often have a reasonable prospect of

[71] Compare Thoma, 2015: 3336, 3348. See the discussion of idealisation in the Interface.
[72] Barry, 1995: 40.

getting one what one wants, they are much cheaper than the actual use of force, and they foreclose fewer options than the use of force does. Therefore, while we should be aware of the possible limitations of a model which assumes that agreement is preferable to conflict for all agents, balancing the various considerations against one another, the model's presumption in favour of agreement seems reasonable. We can then add to these considerations the point I have also made in the cases of the moderate scarcity and interaction conditions: For the contractarian model to be interesting, it suffices that there are some agents who stand to benefit more from agreement, even if this isn't true of all agents.

Finally, is it reasonable to assume that each agent will prefer some form of political order to the absence of any such order? We need to be a little careful in framing this question.[73] It seems likely that there will be some agents who have goals that can only be pursued in the context of political order. These agents, then, will prefer to form such an order rather than live anarchically. If this is so, the question we need to consider is, *given* that there will be some agents creating political institutions, what (prudentially-)ought other agents to do? It is one thing to exist outside of political institutions in a world where there are no such institutions. It is quite another to do so when the world is populated by organised groups of agents reaping productive gains from coordination. In any conflict for resources between agents organised by political institutions and agents not so organised, it is hard to see how the latter could have an advantage.[74] Thus, once some agents have reasons to form political institutions, anarchy – the general absence of political institutions – looks unstable.[75] This doesn't definitively rule out anarchy being preferred by some agents, since I've made a number of further assumptions in arguing that it is unstable, and those conditions might not hold. Nonetheless, they seem plausible. Moreover, as

[73] Widerquist and McCall, 2017.
[74] Compare Rousseau, in *A Discourse on Inequality* (1984: 122).
[75] This gives a contractarian resources with which to answer the 'Mad Max' objection to political society. The objection is that *some* agents could be much better off in the pre-political state than they are in political society – they could be the leaders of powerful biker gangs, for instance. How can *those* agents have greatest prudential reason to enter political society? The answer depends on what one thinks the relevant choice is. Is the alternative to entering political society being the leader of a powerful biker gang in a state of nature? Or is it being imprisoned or killed by the agents of the newly-formed society one has just refused to join? Compare Vanderschraaf, 2019: 205-206.

with the moderate scarcity, interaction, and agreement conditions, the fact that some agents remain outside political institutions still leaves us the question of what those agents who do prefer political order (prudentially-)ought to do. And so the model's fifth assumption is also a reasonable one.

§5.6 Where we are

This chapter has introduced three of the four elements of the contractarian model. I have argued for the coherence and plausibility of a particular conception of rational agents and their preferences, and for a particular characterisation of the hypothetical choice situation in which the agents find themselves.

By way of concluding this discussion of the precontractual conditions, there is one further objection to consider. It goes something like this: The model assumes that its agents exist in a pre-political state of nature. But we don't exist in such a state. However interesting an exercise it might be to work out how such agents would act, doing so can't tell us anything interesting about how we (prudentially-)ought to act, given that we find ourselves in the radically different circumstances of an existing political society. It would be like trying to use our poker model to work out how best to play bridge.

There are two responses to this objection. First, even if the model can't provide us with valid prescriptions for action, because it does not accurately capture the circumstances in which we find ourselves, it is still useful for evaluative purposes. One way of evaluating the political institutions we actually have is to ask if they are the institutions we would choose if we were in a pre-political state of nature, and had a choice. To the extent that the state of nature represents us as free – as not inherently subject to political authority – the contractarian model is a way of testing if our existing institutions are consistent with understanding ourselves as free.

The second response is stronger, however. Why should we concede that the model

does not accurately represent the circumstances in which we find ourselves? It's true that we often think of contractarian models as being grounded in a pre-political state of nature, but none of the five facts about the world I have said the model assumes is inconsistent with reality. Certainly, if the model ultimately recommends political institutions which differ from those we presently have, there are questions about how one might go about getting from where we are to where the model thinks we should be. But the *status quo* is *not* a point from which we *cannot* unilaterally move. The maintenance of the *status quo* requires our continued cooperation. We can threaten to withdraw that cooperation. If enough of us decide to act in concert, we can collapse existing political institutions, and even if we just act individually, we can often worsen the situation of others to some extent, or can credibly threaten to do so. Our political institutions are not fixed by laws of nature or ordained by god. We created them, and we can remake them if we want.[76]

[76] Binmore, 1994: 13-15, 85.

Chapter 6

The final element of the contractarian model is a bargaining theory. A bargaining theory takes as inputs rational agents, their preferences, and their circumstances, and produces as an output an individual rational prescription for each agent. Ultimately, I will argue that, for the relevant set of agents, the content of each prescription will be to establish and maintain liberal political institutions. The main contribution of this chapter will be to establish the conditions under which bargaining theories prescribe an *equal* division of the bargaining surplus. Chapters 7 and 8 will then argue that the contractarian model (approximately) meets these conditions.

In §6.1, I will give a general description of a bargaining problem. In §6.2 and §6.3, I describe three candidate bargaining solutions. One of these – the egalitarian solution – will not be appropriate for use in the contractarian model. In §6.4 and §6.5, I evaluate the remaining two solutions. In §6.6, I state a bargaining theory for the contractarian model.

§6.1 The bargaining problem

Bargaining occurs when one agent, L, proposes to forgo something she wants, α, in order to provide another agent, M, with something he wants, δ, on condition that M agrees to forgo something he wants, γ, in order to provide L with something she wants, β. Rational bargaining occurs when L prefers what she gets to what she forgoes ($\alpha \leq \beta$); and when M similarly prefers what he gets to what he forgoes ($\gamma \leq \delta$). Neither agent gets everything he or she wants – L, for instance, would ideally prefer to get β without giving up α – but each agent is better off under the bargain than under the *status quo ante*.[1]

[1] On the link between bargaining and *modus vivendi* theory, see Gaus (2003: 58-59).

Our most familiar experience of bargaining is buying something.[2] Suppose Liz has money, and Max has coffee. Liz would prefer to have both money and coffee, but she prefers having coffee to having a certain amount of money, $\$\alpha$. Similarly, Max would prefer to have both money and coffee, but he prefers having a certain amount of money, $\$\delta$, to having coffee. So long as $\alpha \geq \delta$ (so long as the dollar value of the coffee to Liz equals or exceeds the dollar value of the coffee to Max), some trade between them is rational.[3]

The key to a theory of rational bargaining is the idea of *concession*. Unless $\alpha = \delta$, there is a range of prices for which Liz would be willing to buy the coffee, and a range of prices at which Max would be willing to sell. The gap between the value Liz places on the coffee and the value Max places on it ($\alpha - \delta$) creates a *surplus* that can be distributed between the two agents.[4] The surplus will only be produced if the agents coordinate. Liz's concession is that part of the total available surplus which she agrees to allow Max to have in return for him coordinating with her to produce the surplus.

A bargaining theory specifies Liz's *rational* concession.[5] This is the part of the surplus which it is rational for Liz to concede to Max. It is the least Liz can concede to Max, assuming he is rational, and still give him sufficient reason to coordinate with her. If she tried to insist on conceding less, it would not be rational for Max to agree to coordinate, and the surplus would not be produced, leaving Liz with the suboptimal *status quo*. If she were to concede more, she would be not be realising her goals to

[2] Binmore, 1994: 186-187.

[3] In this example, the coffee plays the role of both β (what *L* prefers to α) and γ (what *M* gives up to get δ). This is typical of cases involving *trade*, but doesn't hold in general – in bargaining situations, what one person gives up isn't necessarily something another person gains. Note also that, in this example, '$\geq$' can mean 'is (weakly) preferred to' because, for simplicity, I assume the agents to be using the same value scale (dollars). This shared scale is what allows the direct comparison ($\alpha \geq \delta$) between something *L* values (α) and something *M* values (δ). Agents will not in general share such a scale – see §5.1 n1 and n2 for the more general form of preference notation, and §6.2.

[4] This surplus is usually called a *cooperative surplus* (Gauthier, 1986: 141). As noted earlier (§1.5 n37, *supra*), I avoid that label because I want to leave open the possibility that coordination is coerced. Compare Gauthier, 1986: 209; Binmore, 1994: 72.

[5] For the bargaining theories with which I will be concerned, this will be a unique value. Other bargaining theories might instead specify a range.

the greatest degree possible, because Max would have agreed to coordinate at a lower price. Either option would be irrational for Liz. Max's position will be symmetric. He cannot rationally expect Liz to concede more than what it is rational for her to concede; nor would it be rational for him to ask that she concede less. Thus, the division of the surplus represented by Liz's rational concession is also the division to which it is rational for Max to agree. A bargaining theory therefore identifies a point of coordination which is simultaneously rational for each agent.

The agents in the contractarian model are in the same kind of strategic circumstances as Liz and Max.[6] Define S_A (an 'anarchic' state of the world) as a state in which agents do not coordinate their actions with one another, but simply pursue the realisation of their goals by whatever other means they have available. When I say 'by whatever other means', I mean to include anything the agents can do by themselves to realise their goals, including physically forcing other agents to do things.[7] Threats are *not* included here, because in order to be effective threats require the acquiescence of the agent being threatened. That is to say, states of the world realised by threats are *coordinated* states of the world, and S_A is defined as a world in which there is no coordination.

For a given population of agents, S_A will represent a particular distribution of degrees to which agents' goals are realised. This distribution specifies a baseline degree of goal realisation for each agent – this is the degree to which her goals are realised in the absence of any coordination.[8] The baseline represents the *status quo ante* upon which coordinated states of the world must improve if it is to be rational for an agent to agree to produce those states. The degree to which an agent's goals are realised at the baseline may, of course, vary from agent to agent, depending on both the preferences and the capacities of the agents. It is likely that some agents will have

[6] For a summary of the conditions specifying a bargaining situation, see Osborne and Rubinstein (1990: 9-11). The canonical statement is given by Nash (1950: 159; see also Nash, 1953; Binmore, 1998: 46-47, 64).
[7] Binmore, 1994: 25-26; Moehler, 2018: 101.
[8] Compare Gauthier (1986: 130-133), but see also Gauthier's discussion of the baseline and the Lockean proviso (*ibid*.: 190-232). I defend my understanding of the baseline against Gauthier's in §7.5, *infra*.

their goals realised to a greater degree in S_A than others do.[9]

Define S_B (a 'bargained' state of the world) as a state in which the agents do coordinate their actions according to the terms of some agreement. In S_B, each agent achieves a higher degree of goal realisation than she does in S_A. If she did not, she would have no reason to agree to the bargain by which S_B is realised; and if she could be forced to coordinate her actions against her will, she could have been forced in S_A as well, and so the baseline would already reflect that fact.[10] It is unlikely that there is only one way in which the agents can coordinate in moving from S_A. (In the coffee case, if there were only one way to move from S_A, α would equal δ.) As such, there will generally be a set of possible bargained states of the world, $\{S^i_B\}$, with each S^i_B representing a different pattern of coordination which is superior to S_A from the point of view of each agent. Although each agent will prefer each S^i_B to S_A, the agents will probably have different rankings for the different elements of $\{S^i_B\}$. Which is to say, although coordination generates a valuable surplus – an excess of goal realisation which is not present in S_A – the agents are not indifferent about how this surplus is distributed. Assuming it can be distributed in different ways, each agent will want it to be distributed to her own advantage as much as possible.[11]

In analysing bargaining, it is helpful to study a geometric representation of the situation. Figure 1 illustrates a simple world with two agents, *L* and *M*. *L*'s degree of goal realisation is represented on the *x*-axis; *M*'s degree of goal realisation is represented on the *y*-axis. A point in the space (*l*, *m*) specifies a state of the world in which *L*'s goals are realised to a degree *l* and *M*'s goals are realised to a degree *m*. *l*

[9] However, see §7.5.

[10] At this point I am neglecting the possibility of partial coordination. Strictly speaking, S_A is a state of the world without any coordination between agents. If an agent, *L*, cannot be forced to do a particular thing in S_A, this implies that no single other agent is capable of forcing her to do the thing, but it doesn't follow that no group of agents is capable of forcing her, and so it doesn't strictly follow that she cannot be forced to do the thing. The possibility of partial coordination opens up many complications, and will be the main concern of Chapter 8. For the time being, I set it aside. Note that the possibility of partial coordination isn't present in 2-person interactions, because in those cases the only possible configurations of agents are as isolated singletons and as a coordinating pair. So, this problem doesn't arise in the 2-person cases I discuss. See Kalai, 1977: 1623; Binmore, 1994: 75-78.

[11] Rawls, 1999a: 4, 109; Osborne and Rubinstein, 1990: 11.

and *m* can also be called the 'payoffs' received by agents *L* and *M* respectively in a given state of the world.[12]

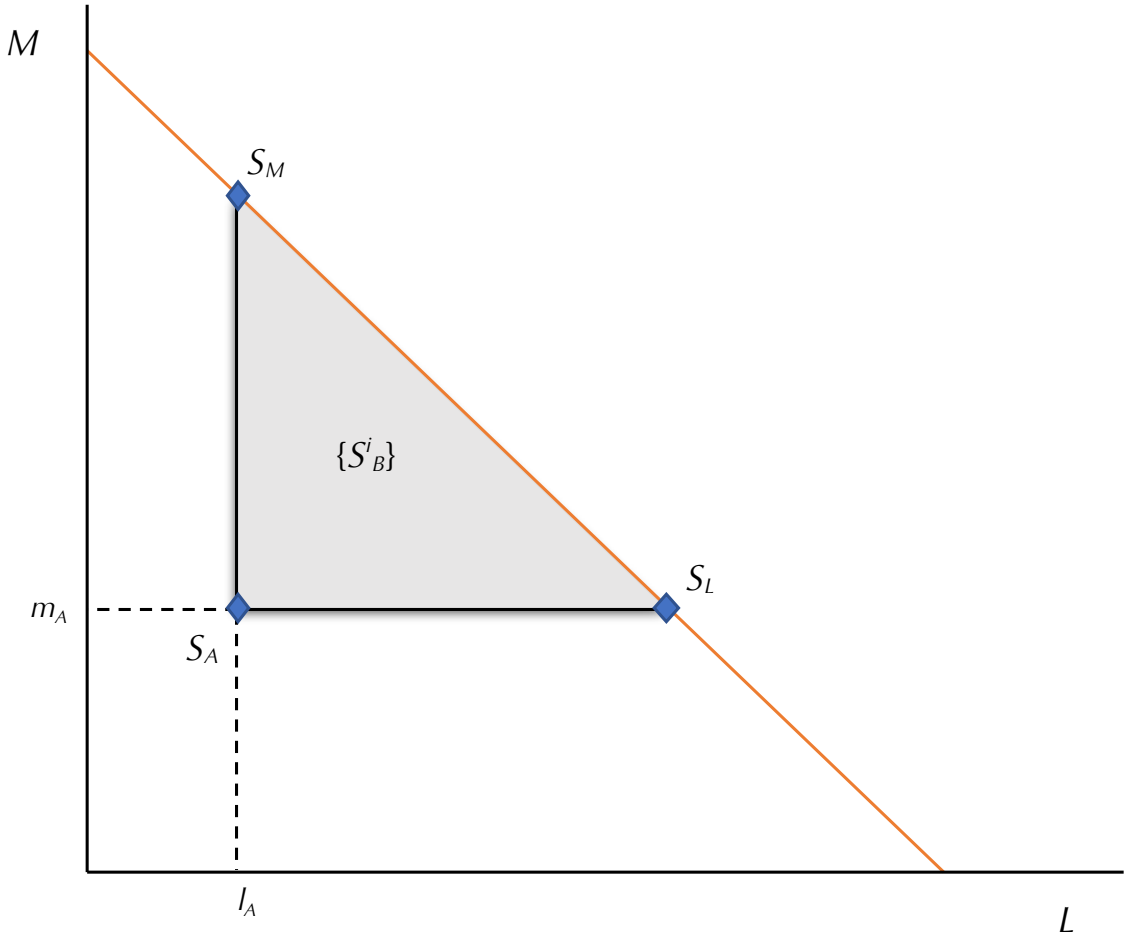


Figure 1

If the agents do not coordinate their actions, they will realise $S_A = (l_A, m_A)$. The shaded area represents $\{S^i_B\}$, the set of possible worlds in which *L* and *M* coordinate their actions such that each prefers the state of affairs realised by their coordinated actions to the state of affairs in which they do not coordinate (for simplicity, I assume that all points in $\{S^i_B\}$ are accessible). The line $S_M S_L$, which includes the line segment $[S_M S_L]$ but extends beyond it, represents the *possibility frontier* for the world. Points

[12] More generally, $\{S^i_B\}$ is an *n*-dimensional space in which each dimension is a measure of the degree to which an agent's goals are realised. A set of coordinates in the space, $S^i_B = (i_1, i_2, i_3 \ldots i_n)$, identifies the degree to which each agent's goals are realised under a particular pattern of coordination, *i*. i_1 is the degree to which Agent 1's goals are realised, i_2 is the degree to which Agent 2's goals are realised, and so on.

to the 'north east' of this line are not accessible to the agents, given the scarcity of resources in the world.[13]

L and *M* have different preference orderings of the points in $\{S^i_B\}$. For instance, *L* prefers a point in the region of S_L to all other points in the space. S_L is a state of the world in which *L* is allocated all of the gains of coordination between the agents. In contrast, *M* ranks some point in the region of S_M above all other points in the space.

A bargaining theory selects one particular element from the available options in $\{S^i_B\}$ – call it S^*_B – and shows that coordinating with other agents to realise S^*_B is what each rational agent to which the theory is applied (prudentially-)ought to do, given the circumstances in which she finds herself. S^*_B is the rationally-mandated compromise between the coordinating agents. Each agent will have a prudential reason to take an action, from her set of available actions, which is consistent with bringing about S^*_B so long as each other agent also takes an action consistent with the realisation S^*_B. The problem, then, is to find a division of goal realisation which each individual agent has a prudential reason to accept and realise.

In the literature on bargaining theory, there are three main strategies for approaching the bargaining problem: the Nash solution, the Kalai-Smorodinsky solution, and the 'egalitarian' solution.[14]

§6.2 The egalitarian bargaining solution

In the egalitarian solution to the bargaining problem, rational agents (in a bargaining situation where each is symmetrically positioned) select S^*_B in such a way as to equalise the utility each agent gains from coordination over and above her utility at S_A.[15] Importantly, 'utility' here does *not* refer to degrees on a vNM utility function.

[13] §5.5.1.
[14] Nash, 1950; Kalai and Smorodinsky, 1975; Kalai, 1977. See also Gaertner and Klemisch-Ahlert, 1991; Vanderschraaf, 2019: 164-166.
[15] Kalai, 1977: 1629. Obviously, not all bargaining situations involve symmetrically positioned

Rather, it refers to something more substantive – classic hedonic utility would be one possibility, but any scale with a similar form could be used.

Insofar as agents have different (hedonic, say) utility functions, this solution will distribute the surplus of goal realisation generated by coordination unequally in order to produce a uniform gain in utility for each agent. Implementing this solution requires, first, that we are able to measure agents' utility levels; and, second, that comparisons between the utility levels of different agents (interpersonal comparisons of utility) are meaningful, such that it makes sense to say that Liz's (gain in) utility is equal to Max's.[16]

Both of these conditions are controversial. On certain understandings of utility there is no obvious mechanism by which we can have access to one another's utility levels; utility is a subjective experience which cannot be measured by anyone other than the agent experiencing it. Quite apart from such epistemic considerations, the idea of interpersonal comparisons of utility has frequently been dismissed as misconceived. It has been argued that Liz's utility and Max's utility are like apples and oranges – they don't exist on a universal utility scale which can tell us that his utility is greater than, less than, or equal to hers.[17]

Much depends here on how 'utility' is defined. There are ways we could define utility which would avoid the difficulties mentioned. We could, for instance, say that utility is something objective and measurable – suppose it consists in an index of nutrition, health, mobility, education, etc., of the kind objective list theories of welfare employ.[18] In principle, there need be no difficulty measuring utility, so understood, or making meaningful comparisons between the utility levels of different agents.

---

agents, but in distinguishing the egalitarian solution from the two standard alternatives it is sufficient to consider the symmetric case. For further discussion, see Binmore (1998: 83-87).

[16] See Moehler, 2018: 52-53. Moehler notes that this is a feature of the bargaining theories employed by, for example, Binmore and Vanderschraaf (and, thus, that their theories will be to that extent moralised).

[17] Compare Binmore, 1994: 50-51; List, 2003.

[18] Dworkin, 1981a: 226.

The problem with such an approach is that, no matter what definition of utility we use, many people will disagree with it. Suppose we grant that, as the egalitarian solution says, rational agents would agree on a bargain which equalises utility gains. Suppose further that we define utility in such a way that resources held by blue-eyed people generate ten times as much utility as resources held by brown-eyed people. Rationality, on this account, would therefore require that blue-eyed people agree to accept 10% of the resources that brown-eyed people get. This definition of utility is absurd, of course, but it draws out an important feature of the egalitarian solution. It only makes sense to say that blue-eyed people would be *irrational* in rejecting the proposed bargain if it is the case that they accept the definition of utility given. If they have a different understanding of utility, then they can rationally reject the proposed bargain, precisely because it fails (in their view) to equalise utility gains.[19] Importantly, *any* objective list conception of utility, no matter how plausible, will have this feature.

So, insofar as the egalitarian solution could be used to make prescriptions to agents in a contractarian choice situation, it must assume that all the agents agree on some value scale. But such an assumption is inappropriate, given the purpose for which the contractarian model is being constructed. One of the basic assumptions I have made about the world is the existence of disagreement about value. There are purposes for which my assumption would be inappropriate – if our purpose were to model a *fair* distribution of the gains from cooperation, it would make sense to assume that the agents all agree on the *correct* conception of value.[20] But that's not

[19] If it turns out that the apparently absurd understanding of utility is actually correct, then the blue-eyed people will be making a *mistake*, but one can make mistakes while acting rationally, if one has imperfect knowledge about the world. Obviously, one could have a thicker, objective conception of rationality, such that one is irrational when one is mistaken about the correct conception of value. But, as the next paragraph explains, this conception of rationality would not be appropriate in the context of a contractarian theory. See also the discussion of idealisation and information in the Interface.

[20] Vanderschraaf, for example, strictly claims that he is discussing problems of fair division (2019: 159ff, 184-188). Perhaps it is not surprising, then, that he endorses the egalitarian bargaining solution. However, the fact that he is a moral contractarian (§1.6 n45, *supra*) consequently creates an ambiguity in his discussion. Are we supposed to understand 'fair' in a revisionist, contractarian way (§1.5 n30, *supra*)? Or does the term retain (traditional) moral content, and thus constitute a success condition for the contractarian account (§1.6, *supra*)?

the kind of purpose to which I propose to put the contractarian model. Idealising agents so that they agree about value would assume away the problem the model is designed to solve.

Therefore, I set the egalitarian solution aside. Even if the problem of interpersonal comparisons is not insuperable, the egalitarian solution is not appropriate for my purpose. This was implicit in §5.2, when I described the von Neumann-Morgenstern method for constructing a cardinal measure of the degree to which agents' goals are realised. Such a measure is predicated on the assumption that direct comparisons of value are unavailable to the agents. The approach taken by the other two bargaining solutions, which I follow, and from which the egalitarian solution departs, is to take agents' preferences, rather than their utility levels, as 'basic'.[21] This approach assumes that each agent has a scale of value – a set of preferences, broadly understood so as to incorporate moral commitments – but that the only comparisons we can validly make between agents are composed of the degrees to which each agent is moving up and down her own scale of value.

§6.3 The Nash and Kalai-Smorodinsky solutions

Having excluded the egalitarian solution, we are left with the Nash and Kalai-Smorodinsky solutions. I'll first state the solutions before going on to evaluate them.[22]

The Nash solution selects S^*_B so as to maximise the function, N:

$$N = (l_B - l_A)^{\lambda} (m_B - m_A)^{\mu}$$

[21] Osborne and Rubinstein, 1990: 11-12.

[22] I follow Binmore (1998: 77-78, 82). It is worth mentioning that both Nash and Kalai and Smorodinsky devise their solutions for bargaining situations with just two participants. For extensions of the solutions, see Osborne and Rubinstein (1990: 23), Imai (1983), and Gaertner and Klemisch-Ahlert (1991: 170). Given that it is easier to get to grips with the solutions in the 2-person case, I will stick with that for the purpose of exposition, but the general shape of the extensions to $n > 2$ will not be difficult to see.

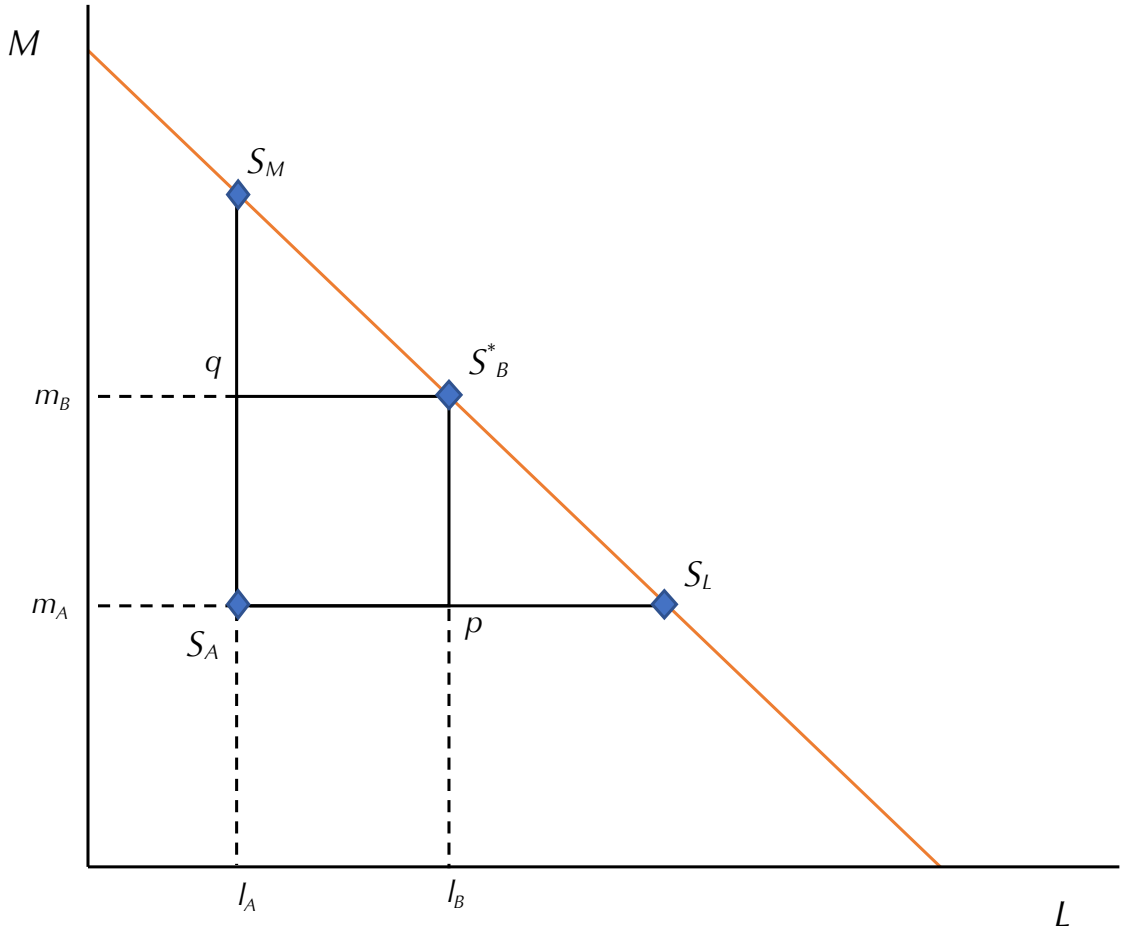


Figure 2

As in Figure 1, l_A and m_A are the payoffs (measured in degree of goal realisation) which *L* and *M* respectively receive when they do not coordinate their actions. λ and μ are exponents representing the agents' 'bargaining power' – *L* has a bargaining power λ and *M* has a bargaining power μ. Bargaining power is a concept to which I will return.[23] For now, it is sufficient to note that it is here constructed as a ratio, λ/μ (the bargaining power of *L* relative to the bargaining power of *M*) such that when the agents have equal bargaining power, we can set $\lambda = \mu = 1$ and neglect the exponents in considering the solution. The only variables remaining are l_B and m_B. S^*_B is defined as the point (l_B, m_B) at which the value of *N* is maximal, subject to the constraint that

[23] See the introductory remarks in Chapter 8, and n4 in particular. One potentially fatal objection to using the Nash solution in the context of a contractarian theory – and, of course, to using any version of the Kalai-Smorodinsky solution which attempts to capture bargaining power in a similar manner – is that agents might disagree about how to characterise bargaining power, in just the way that they could disagree about how to measure utility (§6.2; Thoma, 2015: 3336, 3347).

S^*_B is in $\{S^i_B\}$. For the equal bargaining power case, then, the Nash solution holds that the agents coordinate rationally when they arrange their actions so that the *product* of the *gains* they make from coordination is maximised. A geometric representation of the Nash solution is given in Figure 2.

As can be seen, in the 2-person, equal bargaining power case, implementing the Nash solution amounts to a prescription to each agent to coordinate so as to maximise the area of the rectangle S_A-p-S^*_B-q.

Kalai and Smorodinsky take a different approach. Their bargaining solution selects $S^*_B = (l_B, m_B)$ so as to maximise the lesser of two quantities, K_L and K_M:

$$K_L = \frac{l_B - l_A}{l_C - l_A} \qquad K_M = \frac{m_B - m_A}{m_C - m_A}$$

l_A, l_B, m_A, and m_B are defined as before. The Kalai-Smorodinsky solution introduces a new element, however: the point $S_C = (l_C, m_C)$. l_C is defined as L's maximum possible *claim*. It is the payoff, in terms of goal realisation, which L receives if she is allocated as much of the available surplus as possible. Similarly, m_C is defined as M's maximum possible claim. Typically, it is not possible for both L and M simultaneously to receive their maximum possible claims, and so the claim point, S_C, will not be in $\{S^i_B\}$. It will lie beyond the possibility frontier.

K_L is a term expressing L's gain from coordination, $(l_B - l_A)$, as a fraction of what L could possibly have gained from coordination, $(l_C - l_A)$. Thus, if a bargaining solution allocates L as much of the available surplus as possible (so, $l_B = l_C$), K_L will have a value of 1. If L is allocated no gains from coordination (so, $l_B = l_A$), K_L will have a value of 0. The value of K_L varies between 0 and 1 as L is allocated an increasing proportion of the gains she could possibly make from coordination. In the same way, K_M expresses M's gain from coordination as a fraction of what he could have gained. The Kalai-Smorodinsky solution holds that agents coordinate rationally when they arrange their actions so that the agent who is gaining a lesser proportion of what he

or she could have gained, is gaining as high a proportion of what he or she could have gained as possible.[24]

As before, it can be helpful to look at a geometric representation of the bargaining problem to see how the proposed solution works. Figure 3 shows the Kalai-Smorodinsky solution. The coordination point, S^*_B, is defined as the point of intersection between the possibility frontier and the line joining the non-coordination point and the claim point.

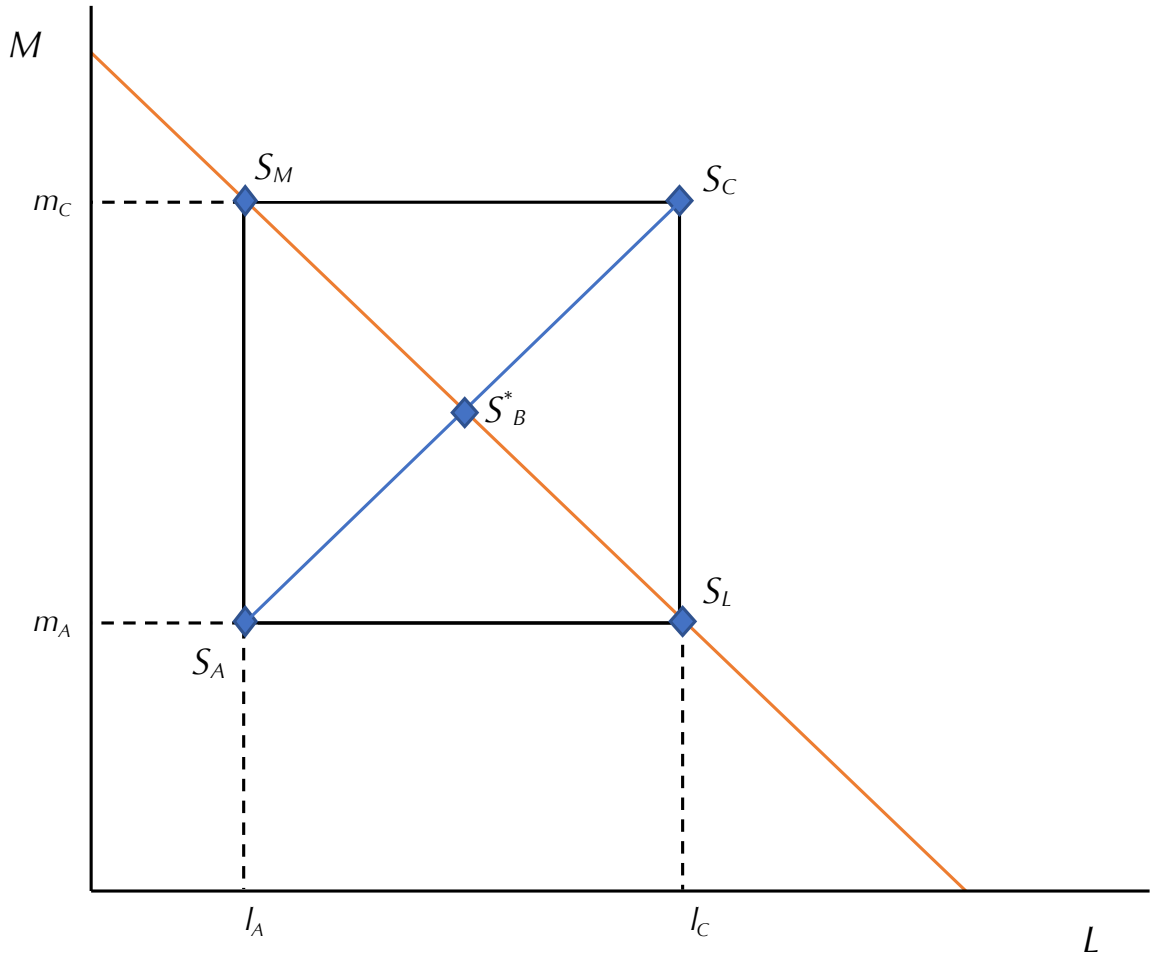


Figure 3

Statements of the two solutions as mathematical abstractions offer a poor perspective

[24] Note that the sum of K_L and K_M need not be 1. When the possibility frontier is a straight line, $K_L + K_M = 1$. However, when the frontier is (strictly) convex, $K_L + K_M > 1$. At the extreme, when it is actually possible to reach the claim point, S_C, then $K_L + K_M = 2$.

from which to evaluate their merits. As I have set out above, for my purposes, a bargaining theory is a tool designed to solve a problem that agents in a contractarian model face. The agents can all gain from coordination with one another, but they have a variety of coordination options, and they rank these options differently. A bargaining theory is successful if it identifies a coordination option to which each agent simultaneously has greatest prudential reason to agree. So, in evaluating the Nash and Kalai-Smorodinsky solutions, we need to ask which solution tracks agents' prudential reasons.

Broadly speaking, there are two ways of approaching this question. The first option is to take what I will call a *procedural* approach.[25] On this kind of approach, we imagine ourselves in the position of agents bargaining over some prize and try to model how the back and forth of negotiation would proceed with a view to identifying a unique point at which rational agents would agree to compromise. Insofar as this process picks out a uniquely rational point of compromise, the bargaining problem will be solved.

One problem with taking a procedural approach is that one needs to introduce various constraints on the negotiation model in order to narrow the field of possible coordination points down to a single point. It is one thing to justify these constraints for some specific real-world case (wage bargaining between an employer and an employee, for instance). It is more difficult to justify them for something as abstract as a contractarian choice situation. Since the constraints will essentially determine the prescription of the model, constraints which are *ad hoc* will give rise to an *ad hoc* prescription.

Because they are interested in a general solution to the bargaining problem, both Nash and Kalai and Smorodinsky instead take an *axiomatic* approach. They set out to find the most parsimonious and intuitively plausible set of constraints with which to characterise a bargaining solution that will nonetheless exclude all but one of the

[25] Rubinstein (1982: 98) calls this a 'strategic' approach.

possible forms of coordination available in a generalised construction of the bargaining problem.[26] For example, as I'll discuss below, each of the two solutions insists that a coordination point must be Pareto-efficient, on the basis that no rational agent could have greatest prudential reason to agree to coordination at a particular point if coordination at some other point would be an improvement for everyone. If rational bargains must conform to a few key conditions like that, and if it turns out that conformity with those conditions identifies a unique coordination point, then the bargaining problem will be solved.

Ideally, we don't want the procedural and axiomatic approaches to be alternatives so much as complements for one another. If we have a procedural story we find persuasive, we will want to be able to axiomatise it; and if we have a set of axioms we find attractive, we will want to be able to connect them to a plausible account of how rational agents would behave in an actual bargaining situation.[27]

§6.4 A procedural comparison between Nash and Kalai-Smorodinsky

To make things more concrete, consider the following example from Gauthier.[28] Sam McGee, a prospector, "discovers the richest vein of gold in the Yukon, but lacks the necessary cash (say $100) to register a claim to it."[29] Grasp, a banker, is the only person with cash available to lend to McGee. Let the value of the claim be *V*; let this also be the full value of the surplus available from coordination in the circumstances, and let McGee and Grasp both have vNM utility functions which are straight lines, increasing with increases in the dollar values of their shares of the claim. Suppose

[26] For a critique of this approach, based on the claim that it illicitly *assumes* the existence of a unique solution, rather than arguing for one, see Thrasher (2014: 695). Such a critique is misplaced. So long as the various axioms are themselves independently defensible, there is no circularity.
[27] See Nash, 1953: 129; Binmore, 1998: 42-49; Osborne: 2004: 465-490; Thoma, 2015: 3338; Vanderschraaf, 2019: 163-167.
[28] 1986: 153-154. For Gauthier's description of his bargaining theory, see *ibid*.: 129-146. For criticism, see Binmore (1993: 141-155). Gauthier's bargaining theory is equivalent to the Kalai-Smorodinsky solution for the 2-person case (Gauthier, 1986: 130 n14). For the extension, see Gaertner and Klemisch-Ahlert (1991: 170-175).
[29] Gauthier, 1986: 153.

that each will receive nothing in the event of a failure to strike a deal. Assuming McGee and Grasp are rational agents, what bargain (prudentially-)ought they strike?

First, consider what the Kalai-Smorodinsky solution requires. Let C_{McGee} be the value of McGee's share of the claim. Given that McGee receives \$0 in the absence of an agreement, and that his maximum share of the surplus is V, then K_{McGee} is:

$$K_{McGee} = \frac{C_{McGee} - 0}{V - 0}$$

Similarly, if C_{Grasp} is the value of Grasp's share of the claim, then K_{Grasp} is:

$$K_{Grasp} = \frac{C_{Grasp} - 0}{V - 0}$$

The Kalai-Smorodinsky solution requires that we maximise the lesser of K_{McGee} and K_{Grasp}. In these particular circumstances, this means that we need to equalise K_{McGee} and K_{Grasp}. If one is dividing a fixed sum, V, and if one can freely transfer parts of V between the shares, then the lower of the two shares will be maximised when the shares are equally divided.[30] Since the only variables in the K-terms are C_{McGee} and C_{Grasp}, equalising K_{McGee} and K_{Grasp} requires equalising C_{McGee} and C_{Grasp}. Thus, the Kalai-Smorodinsky solution requires that McGee and Grasp agree to split the value of the claim equally.

Gauthier has a procedural story to accompany this example.[31] Suppose that McGee initially offers Grasp a 20% stake in his claim. As a rational agent, should Grasp accept? Gauthier says no. He reasons that, as the proposal stands, the cost to McGee of failing to reach an agreement is 80% of the value of the claim, whereas the cost to Grasp is just 20%. Since McGee has more to lose from a breakdown in

[30] The assumption upon which this result depends – what is known as 'transferrable utility' or 'transferrable payoff' – is not trivial. It will hold in some cases, but not in others (Binmore, 1994: 260, 282; Binmore, 1998: 38-39, 64, 174-175; Osborne, 2004: 241).
[31] 1986: 73-75, 136-139, 146-150; compare Zeuthen, 1930: 104-121.

negotiations, he is in a weaker bargaining position, and Grasp can exploit this by holding out for a better deal. This is true so long as the proposal on the table allocates a greater share of the claim to McGee than to Grasp; and, of course, the reverse is true if the proposal is to allocate more than half of the claim to Grasp. Only an equal division of the claim leaves neither party with reason to think that the other has more to lose by prolonging a deadlock.[32] The Kalai-Smorodinsky solution formalises this procedural story. It requires that the agents coordinate in such a way as to ensure that the agent who has least to lose from prolonging a deadlock (because her K is lower than any other agent's) is getting as great a payoff as possible from the agreement.

Next, consider what the Nash solution requires. Implementing the Nash solution requires that agents maximise the product of their payoffs, adjusting for differences in bargaining power where appropriate. If we assume, for the sake of argument, that McGee and Grasp have equal bargaining power, and that, as before, McGee and Grasp both get nothing in the event that no deal is reached, implementing the Nash solution requires that we maximise the product of the share of the claim allocated to McGee (C_{McGee}) and the share allocated to Grasp (C_{Grasp}). By inspection, we can quickly see that this product will be maximised when the claim is divided equally.[33]

[32] Suppose the following objection is raised. Suppose Grasp insists on 80% of the claim, and simply refuses to budge on the matter. Grasp is a wealthy banker; he can afford to wait McGee out, refusing to compromise while living comfortably. McGee, on the other hand, can't afford to delay. If he doesn't settle the claim, he won't be able to buy food. It is important to realise that the details that have just been added to the example have altered it in a substantive way. If Grasp's coordination is needed in order for McGee to avoid starving, then avoiding starvation is a component of the surplus over which the agents are bargaining. Given the additional details, the bargaining position of the parties has changed, and the bargaining outcome will also change. Gauthier's bargaining solution will still recommend an equal division of the surplus, but the surplus is now the vein of gold plus some other things. If we assume that McGee puts a considerable premium on not starving, it is plausibly the case that not starving plus 20% of the claim represents half the available surplus. Thus, it would be a rational bargain. Note further that although the example itself is set up in such a way that time passes between 'rounds' of offers and counter-offers, this is just for ease of explanation – the example is supposed to model idealised rational agents who will immediately recognise the rationally prescribed bargaining solution (compare Binmore, 1994: 57, especially n68). Compare also Hobbes's discussion of the ninth and tenth Laws of Nature (1996: 107-108).

[33] Suppose we split V equally between McGee and Grasp, and each receives c. The product of their payoffs is then c^2. Suppose next that we allocate slightly less to McGee ($c - 1$) and slightly more to Grasp ($c + 1$). The product of their payoffs is then $c^2 - 1$. Clearly, this is less than c^2. If we allocate less again to McGee ($c - 2$) and more to Grasp ($c + 2$), the product of their payoffs will be $c^2 - 4$, and so on. Obviously, we can transpose the labels on the agents and it will make no difference. Thus, when an equal division is possible, the Nash solution will select it.

Telling a procedural story to accompany the Nash solution requires a little more work, but the example can be modelled as a form of alternating offers game.[34] Suppose that McGee gets to make a proposal for a division of the claim, and that Grasp can then either accept or reject the offer. If Grasp accepts the offer, the claim is divided according to the proposal, and the game ends. If Grasp rejects the offer, a period of time passes, and then he gets to make an alternative offer. McGee can either accept this offer, in which case the game ends, or he can reject it, wait some period of time, and make a further offer, and so on. The catch is that we assume the agents to discount the future in such a way that they always prefer a dollar now to a dollar at some later point. Thus, rejecting an offer, and allowing a period of time to pass, is costly for the agents – it effectively shrinks the size of *V*.

Clearly, the agents have an incentive to make a deal immediately, rather than wasting some part of the value of the surplus haggling. For simplicity, assume the agents are very impatient, such that (for each agent) one third of the value of the claim will be exhausted after one negotiating period, and that the entire value of the claim will be exhausted (for each agent) after two periods.[35] If Grasp is making the second proposal, he knows that McGee will accept anything he proposes, since rejecting the offer will leave McGee with nothing. Thus, in the second negotiating period, Grasp can propose to keep the entirety of the remaining cooperative surplus (less some arbitrarily small ε). This will amount to (nearly) two thirds of the initial claim. Knowing that he can guarantee himself (nearly) two thirds of the claim simply by waiting, Grasp will refuse to accept any offer less than two thirds of the surplus in the first negotiating period. Since McGee also knows this, he will propose to offer Grasp two thirds in the first negotiating period, and to keep one third for himself. Since Grasp can't do any better than this by waiting, he will accept McGee's offer.

[34] Binmore, 1998: 122-130; see also Rubinstein, 1982; Hargreaves Heap and Varoufakis, 1995: 128-137.

[35] Additional complications arise when the agents have different 'time-preferences'. If one agent is more patient than another, the value of the claim for the more patient agent shrinks more slowly. This measure of an agent's patience is modelled in Rubinstein's alternating offers game as a difference in the agents' relative bargaining powers (λ and μ) as those are represented in the Nash bargaining function, N. See Binmore, 1998: 70-73, 121-122, 126-128; Thoma, 2015: 3348.

This brief sketch is intended only to give a vague sense of how a procedural story for the Nash solution can work. Filling in the details, we can show that if we allow the time period between offers to approach zero, and model the negotiations as an infinite game in which each agent has a constant rate for discounting future gains, the alternating offers game has a unique equilibrium which corresponds to the general form Nash bargaining solution.[36] If we allow McGee and Grasp to be equally patient, we get the straightforward result that they (prudentially-)ought to split the claim equally.

I want to take several points away from this procedural discussion of the two bargaining solutions. First, it speaks strongly in favour of the Nash solution that it can be derived from a non-cooperative game of the form of the alternating offers game.[37] Game theoretic solutions to such games are paradigmatic examples of agents acting on the basis of their prudential reasons. They provide a much firmer foundation than the kinds of plausible but informal psychological stories we can tell about agents anticipating one another's fear of deadlock in cases like the McGee-Grasp scenario.[38] Gauthier himself at one point rejected the Kalai-Smorodinsky solution in favour of the Nash solution for this reason.[39] That said, the fact that Gauthier does not express his original procedural understanding in so formal a manner does not imply that no such formal statement can be given.[40] In §6.5, I will register some doubt that the reasons we have to accept the Nash solution are decisive.

---

[36] Binmore, 1998: 122-130.
[37] Moehler, 2018: 55-56, 108.
[38] However, see Hargreaves Heap and Varoufakis (1995: 120, 122), Thoma (2015: 3336, 3343, 3347), and Vanderschraaf (2019: 166-167). Vanderschraaf notes that the Rubinstein approach only guarantees determinate solutions given highly restrictive conditions (an infinitely divisible good, no more than two players, no significant time preferences (see n35, *supra*)). Thoma argues that if we assume (i) that bargaining is costless and (ii) that the process of bargaining itself has to be agreed by the agents, then there is *no* standard non-cooperative game which can be used to model bargaining. I have proposed accepting (i) in §5.5.2, *supra*. I take it the argument of §7.5, *infra*, strongly indicates (ii), even if it is not strictly entailed.
[39] 1993: 177-181. He has since reversed himself (2013: 609-614).
[40] See Binmore, 1998: 93; Moulin, 1984; Thoma, 2015: 3343 n4.

The second thing to note is that, given certain conditions, it doesn't matter which solution we use, because the solutions converge on the same coordination point (albeit for somewhat different reasons). The key factor determining this convergence turns out to be a property of the solution space which, up to this point, I have taken for granted. In the examples I have considered, the solution space is *symmetric*. This means that if coordination on a particular point (*l*, *m*) has been possible, coordination on the point (*m*, *l*) has also been possible; more prosaically, it means that if an 80-20 split of some prize has been possible, a 20-80 split has also been possible. In symmetric bargaining problems, the Nash and Kalai-Smorodinsky solutions select the same coordination point.[41] Although the solutions seldom come apart to a great extent even in asymmetric cases, the fact that they diverge to some degree in those cases means a contractarian model cannot simply avoid a decision between them.[42]

Finally, note that, given certain conditions, both bargaining solutions recommend an *equal* division of the gains of coordination.[43] My claim will ultimately be that, allowing for a few complications, these conditions hold in the contractarian model, and thus the agents in the model have greatest prudential reason to divide the gains of their coordination equally. They do this, I will argue, by selecting liberal political institutions to govern their interactions with one another.

The egalitarian character of the bargaining solutions depends on a number of things. To begin with, it is not at all surprising that a sensible bargaining solution would prescribe egalitarian coordination in a case where the agents and their positions are perfectly identical. Any reason for rational agents to realise a coordination point which advantages one of them at the expense of some other would have to be based on some difference, either between the agents themselves, or in their circumstances.

[41] Binmore, 1994: 81-82; Binmore, 1998: 142.
[42] Binmore, 1994: 82. I follow Binmore (*ibid*.: 82-83; 1998: 142) in thinking that the contractarian choice situation might be asymmetric in this sense. See §7.6, *infra*.
[43] Note the contrast with the egalitarian bargaining solution (§6.2). The egalitarian bargaining solution distributes the surplus *unequally* in order to achieve equal gains in (for example, hedonic) utility. The Nash and Kalai-Smorodinsky solutions (given the relevant conditions) distribute the *surplus* equally, which will not typically result in equal gains in utility.

If there is no such difference, an egalitarian result follows immediately, simply on the basis of a principle of insufficient reason.

Including symmetry in the solution space, which I have just mentioned, there are five key ways in which the agents or their circumstances might be similar or different:

> The agents in the model are *equally rational* if each reasons in the same way.
>
> The agents in the model either have *equal information* or they do not.
>
> The agents are *equally endowed* if their goals are realised to the same degree at the non-coordination point, S_A.
>
> The agents have *equal prospects* if the solution space, $\{S^i_B\}$, is symmetric.
>
> The agents either have *equal bargaining power* or they do not.

Of these, only the first and second are assumed by the way in which I have set up the contractarian model. However, for the sake of simplicity, the examples I have used to illustrate the bargaining solutions have also assumed the other three. Clearly, the model cannot just take this degree of similarity for granted. As such, I discuss the equal rationality, information, endowment, and prospects of agents in Chapter 7. I discuss agents' bargaining power in Chapter 8.

§6.5 An axiomatic comparison between Nash and Kalai-Smorodinsky

The alternative to a procedural approach is an axiomatic approach. In their classic formulations, each of the Nash and Kalai-Smorodinsky solutions is characterised by four axioms.[44] They share three of these: invariance, symmetry, and efficiency.

[44] Plus a number of technical premises about $\{S^i_B\}$. See Binmore, 1998: 60-64, 89-95; also Nash, 1950; Nash, 1953: 136-140; Kalai and Smorodinsky, 1975; Osborne and Rubinstein, 1990: 11-13,

Invariance is a condition which requires that the coordination point (S^*_B) selected by the bargaining solution does not change if we alter the value scales being used by the agents in the model.[45] Solutions which satisfy invariance do not require making interpersonal comparisons of utility between the agents (though they do, of course, rely on *intra*personal comparisons – they require that agents' have vNM utility functions of the kind described in §5.2).[46] Insofar as we think making interpersonal comparisons of utility is problematic, it is reasonable to require that bargaining solutions satisfy the invariance condition. Invariance is the axiom which the egalitarian solution rejects.[47]

I have already noted the *symmetry* condition.[48] It requires that, if a point (l, m) is in $\{S^i_B\}$, then (m, l) is also in $\{S^i_B\}$.[49]

Finally, *efficiency* is a Pareto optimality condition. It requires that the solution fully exhaust any surplus that coordination generates. Referring to Figure 1 again, efficiency requires that S^*_B lie somewhere on the line $S_M S_L$. Other points are either inaccessible (due to scarcity) or represent states of the world in which the position of at least one agent can be improved without making the other agent worse off. The thought underlying the efficiency axiom is that rational agents would neither try to

21-22; Gaertner and Klemisch-Ahlert, 1991; Hargreaves Heap and Varoufakis, 1995: 122-128; Thoma, 2015: 3339.

[45] Within certain limits – technically, the solution must be invariant up to positive affine transformations of the agents' vNM utility functions. See Nash, 1950: 156-157; Binmore, 1998: 89-91. Such transformations preserve the 'shape' of an agent's utility function while allowing it to vary in its slope (so long as it keeps the same sign) and in its position within the solution space. Thus, a utility function can be compressed or stretched, but the transformation will preserve its ordering of options and the ratio of 'distances' between the options. See Binmore, 1994: 53-54 n64.

[46] Thoma, 2015: 3340.

[47] §6.2. In part because he thinks worries about interpersonal comparisons of utility are overblown, Vanderschraaf rejects invariance and defends the egalitarian solution (2019: 185-187).

[48] §6.4.

[49] As also noted, there are versions of both the Nash and the Kalai-Smorodinsky solutions which dispense with the symmetry axiom, but the initial formulations of each included it. Nash's original discussion of symmetry (1950) has caused some confusion (Thrasher, 2014; Thoma, 2015: 3343, 3353; Moehler, 2018: 128-129; Vanderschraaf, 2019: 220 n75, 221 n76, 318 n75; see §1.6.3, *supra*). His formal statement of the condition is as I have described it – a property of the solution space. However, in his informal discussion, he represents the condition as the equal rationality assumption. See Binmore, 1998: 78-79, 91.

coordinate on the impossible, nor leave any part of the surplus unclaimed.

The Nash and Kalai-Smorodinsky solutions differ in their fourth axiom. Nash requires that the solution satisfy an axiom he calls 'independence of irrelevant alternatives'. Kalai and Smorodinsky replace Nash's independence axiom with a monotonicity axiom.[50] The Nash and Kalai-Smorodinsky theorems prove that, with the addition of the fourth axiom, each bargaining solution selects a uniquely rational coordination point in the solution space.[51] Our question is, which of the axioms is it more appropriate for the contractarian model to employ?

What Nash means by his *independence* condition is contraction/expansion consistency.[52] Take the set $\{S^i_B\}$ and let p be the point in $\{S^i_B\}$ selected as the coordination point by a bargaining solution. Now consider two more sets. First, $\{R^i_B\}$, a subset of $\{S^i_B\}$ created by contracting the possibility frontier so that the set of possible bargains is smaller. Second, $\{T^i_B\}$, a set of which $\{S^i_B\}$ is a subset; $\{T^i_B\}$ is created by expanding the possibility frontier so that the set of possible bargains is larger. Nash's independence condition now requires two things. First, if p is an element of $\{R^i_B\}$, then p must be the coordination point selected from $\{R^i_B\}$ by the bargaining solution. That is, shrinking the set of possible bargains must not change the coordination point selected unless the set shrinks in such a way that the previous coordination point becomes unavailable. This is contraction consistency. Second, if p is *not* the coordination point selected from $\{T^i_B\}$ by the bargaining solution, then the coordination point cannot be any other member of $\{S^i_B\}$ – it must be one of the newly available points in $\{T^i_B\}$. That is, enlarging the set of possible bargains can only change the coordination point if the new coordination point wasn't available before. This is expansion consistency. Taking the two requirements together, contraction/expansion consistency requires that changing the size of the set of possible bargains doesn't alter the ordering a bargaining solution gives to the initial

[50] Once again, it is worth emphasising that these independence and monotonicity axioms are not the same as the monotonicity (or independence, or sure-thing) condition I introduced in §5.2, despite the similar terminology. The latter is a constraint on an agent's preference set. The former are candidate constraints on a bargaining solution.
[51] Nash, 1950; Kalai and Smorodinsky, 1975.
[52] Nash, 1950: 159; Nash, 1953: 137, 138; Gaertner and Klemisch-Ahlert, 1991: 165.

set of points, $\{S^i_B\}$.

To put the requirements in less formal terms, contraction consistency says that if Holmes is the best detective in England, then he must also be the best detective in London (the set of detectives in London being a subset of the set of detectives in England). Expansion consistency says that, if Holmes is the best detective in London, but is *not* the best detective in England, the best detective in England cannot be Watson (because, if Watson were a detective, and if he were the best detective in England, he would also be the best detective in London; by hypothesis, Watson is not the best detective in London, and so he cannot be the best detective in England).

Contraction/expansion consistency looks like a perfectly natural requirement when we're talking about a predicate like 'best detective'. But consider the following claim: If Tony is the most popular person in Britain, he must also be the most popular person in Bristol. This is clearly false – Tony might be wildly popular in Britain, despite being highly unpopular in Bristol. Similarly, that Tony is the most popular person in Bristol, but not the most popular person in Britain, does not imply that no other person in Bristol can be the most popular person in Britain. 'Most popular person' is not a predicate with respect to which contraction/expansion consistency is required. So, contraction/expansion consistency is a reasonable requirement in some cases, but not in others. Nash thinks it is reasonable to require that bargaining solutions satisfy contraction/expansion consistency. But this is an open question.

In the context of individual choice, Nash's independence condition is clearly defensible. Suppose Sid is ordering ice cream, and is choosing from chocolate, strawberry, and vanilla flavours. Suppose that, given this set of options, he selects chocolate. Suppose we then tell Sid that we've run out of vanilla. Suppose Sid were to reply that, seeing as there's no vanilla, he will elect to have strawberry. Absent some special explanation (perhaps he has promised always to order strawberry when faced with a choice between chocolate and strawberry) we would think that something very strange has happened. The presence or absence of an apparently irrelevant alternative – vanilla – seems to be affecting the ordering of Sid's

preferences over chocolate and strawberry. We can conclude that Sid is irrational.[53]

We might then generalise this to instances of social choice.[54] Suppose Sid and Tony are splitting a portion of ice cream, and must decide which flavour to order. As before, suppose they settle on chocolate, and we then announce that we've run out of vanilla. Again, doesn't it seem strange that this announcement might prompt them to switch from chocolate to strawberry? Something like this thought – the idea that social choice should be held to the same standards of consistency as individual choice – motivates the independence condition. This has been a particularly influential idea in the literature on democratic theory, where people are often committed to the claim that democratic decisions express the will of a collective agent called 'the people'. Insofar as we want to conceive of a democratic people as a singular, coherent, responsible agent, it seems to make sense to hold that agent to the standards of rationality we apply to ordinary individuals.[55]

Nonetheless, we need to be careful about moving too quickly from standards for individual choice to standards for social choice. Suppose Sid and Tony are planning a trip to the cinema. There are four films showing. Sid ranks the films as follows: $F_1 > F_2 > F_3 > F_4$. Tony, on the other hand, has the opposite preferences: $F_4 > F_3 > F_2 > F_1$. Given these options, suppose Sid and Tony settle on F_3. Neither is getting exactly what he wants, but they compromise in the interests of seeing a film together. Suppose that, when they get to the cinema, they discover that F_4 has sold out, and so they couldn't have seen it in any event. Does it obviously follow that it is still rational for them both to agree to see F_3? The Nash position holds that it is, because F_4 is an irrelevant alternative – its presence or absence cannot alter the social ordering of F_1, F_2, and F_3, and F_3 ranked above the other options in the initial ordering. But now it seems that Sid might have a reasonable complaint. He agreed to see F_3 when doing so was a compromise between his preferred option and Tony's preferred option. Given the new option set, however, F_3 doesn't represent a

[53] For some further discussion, see Thrasher (2019: 438-441).
[54] Binmore, 1998: 94-95.
[55] Compare *ibid*.: 96.

compromise at all – Tony is getting his highest ranked available option, and Sid is getting his lowest ranked available option. If Sid were to propose that they compromise on F_2 instead, it is hard to say how he would be acting irrationally, even though his proposal violates contraction consistency. Once we step away from the quasi-individual, democratic understanding of social choice, and instead see social choice as an exercise in compromise – in conflict, concession, and bargaining – it no longer seems strange that a rational social choice rule would exhibit characteristics that differ from those we observe in individual rational choice.

The *monotonicity* condition introduced by Kalai and Smorodinsky captures this intuition.[56] In a bargaining solution, monotonicity requires that if the set of available options $\{S^i_B\}$ is expanded to $\{T^i_B\}$ in such a way that for any payoff level received by *L*, the maximum possible payoff level received by *M* increases, the solution must select a coordination point in $\{T^i_B\}$ which allocates to *M* a payoff at least as great as the payoff allocated to *M* by the solution in $\{S^i_B\}$. Less formally, monotonicity requires that, if circumstances change such that an agent, *M*, can convert resources to goal realisation more efficiently, the fact that he can do so must not disadvantage him. Replacing Nash's independence axiom with monotonicity yields the Kalai-Smorodinsky bargaining solution.

Qualitative comparisons of the independence and monotonicity axioms are rare in the literature.[57] Formal theorists such as Nash, Kalai, and Smorodinsky, and those who follow them, focus their attention on mathematical rigour. They devote relatively little attention to explaining the motivation for their axioms. On the other side, political philosophers are seldom interested in the mathematical details of an account of bargaining, and so also avoid qualitative analysis of these axioms. In my discussion, above, I have drawn on the work of Ken Binmore, as one of the few theorists bringing both mathematical and philosophical expertise to bear on the

[56] Kalai and Smorodinsky, 1975: 515; Binmore, 1998: 91-93.
[57] *Ibid*.: 92. Muldoon (2016: §4.6), Moehler (2018: 54-56, 85, 128), and Vanderschraaf (2019: 164-167, 185-186) each discuss the issue briefly. Muldoon is not strongly committed on the question. Moehler and Vanderschraaf's arguments are similar to (though somewhat less thorough than) Binmore's, which I discuss and reject in the text.

question. I do so again now.

Binmore argues in favour of the Nash solution – and hence the independence axiom – against the Kalai-Smorodinsky solution and monotonicity. He discusses two positive arguments in favour of Nash and independence, one intuitive and one procedural; similarly, he gives two negative arguments against Kalai-Smorodinsky and monotonicity, one intuitive and one procedural.

The positive intuitive argument for the independence axiom which Binmore describes is a version of the ice cream problem I have outlined above.[58] I have already suggested, and Binmore recognises, that this argument isn't decisive.[59] The cinema trip example is designed to illustrate the weakness of the intuitive argument for independence.

I have briefly discussed Binmore's procedural arguments in §6.4. To summarise, these amount to the claim that the Nash solution can be characterised as a form of non-cooperative alternating offers game (the positive argument), while no such characterisation has been produced for the Kalai-Smorodinsky solution (the negative argument).[60] In combination, these are a persuasive argument in favour of Nash and independence – persuasive enough to at one point convince Gauthier to abandon Kalai-Smorodinsky – but the absence of a non-cooperative characterisation of Kalai-Smorodinsky is not evidence that no such characterisation can be produced.[61]

It remains to examine Binmore's negative intuitive argument. His discussion is short, so I will quote it almost in its entirety.[62] Binmore begins by noting that, on the Kalai-

[58] Binmore, 1998: 94-95.
[59] *Ibid*.: 96.
[60] *Ibid*.: 93, 96.
[61] See n38 and n40. Further, in his most recent work, Gauthier (2013: 609-610) implies that the procedural argument which at one point convinced him of the superiority of the Nash solution amounts to begging the question against (his version of) Kalai-Smorodinsky. See also a suggestive remark by Nash (1953: 136) noting that procedural approaches are hostage to the precise characterisation of the bargaining situation which is given, in a way that axiomatic approaches are not (on this problem, see Vanderschraaf, 2019: 166-167). This speaks in favour of a certain priority for an axiomatic approach.
[62] Binmore, 1998: 92-93; see also *ibid*.: 69-70.

Smorodinsky solution, whenever an agent's risk aversion increases, her bargaining position improves. He illustrates this property of the solution using a divide-the-dollar game analogous to Gauthier's McGee-Grasp scenario.[63] He then claims that this is an implausible implication for a bargaining theory to have:

> "But is this a sensible property for a bargaining solution? Why should it be a strategic asset to a player that he has become less willing to take risks? An increase in his risk-aversion will presumably mean that some of the strategic ploys that seemed attractive before will cease to be tempting because of the high level of risk they involve. Handicapping Eve in this way ought to be to Adam's advantage rather than Eve's. That is to say, an increase in Eve's risk-aversion should benefit *Adam* rather than Eve – as indeed it does in the case of the Nash bargaining solution.
>
> The story which accompanies Individual Monotonicity therefore has only a surface plausibility. In particular, it fails to hold water in the case of dividing the dollar. It remains conceivable that increasing an agent's risk-aversion may be advantageous in other bargaining problems, but it is enough to produce one counterexample to demolish a general claim."[64]

He later adds:

> "[When] Eve becomes more risk averse, the Nash bargaining solution awards her less money than she got before, whereas the Kalai-Smorodinsky solution awards her more money than before. [My earlier discussion of this] comes down firmly in favour of the former being the more plausible of the two possibilities."[65]

Unfortunately, this argument rests on a mistake. It is true that, on the Kalai-

[63] §6.4.
[64] Binmore, 1998: 93, emphasis in original. In his 2-player examples, Binmore's agents are 'Adam' and 'Eve'.
[65] *Ibid*.: 95.

Smorodinsky solution, when a player becomes more risk averse, there is a sense in which her bargaining position improves. Crucially, however, in the divide the dollar game which Binmore discusses, Eve's position improves in *vNM utility* terms, not in *monetary* terms.[66] That is, when Eve becomes more risk averse, the Kalai-Smorodinsky solution allocates to her a higher degree of goal realisation than before. It does *not* give her more money. It would certainly be strange if becoming more risk averse in the divide-the-dollar game allowed an agent to gain a greater share of the dollar than before. But that isn't what happens.

To see why, we need to examine what risk aversion means in the context of a bargaining situation. When both agents are risk neutral, a geometric representation of the situation will look something like Figure 1 – in particular, the possibility frontier will be a straight line, indicating that a unit loss of goal realisation for one agent will be a unit gain for the other, and that the loss of a unit of goal realisation matters the same to an agent whether it happens near his or her ideal point or near the other agent's ideal point. When an agent is risk averse, this isn't the case. Risk averse agents care more about losses and gains when they have very low degrees of goal realisation, and care less when they already have a high degree of goal realisation. That is to say, in the divide the dollar game, they prioritise the first ten cents they get over the second ten cents, and the second ten cents over the third. They care much less about the last ten cents of the dollar they could receive than they do about the first ten cents. If the risk averse agent is bargaining with a risk neutral agent, it is then no longer the case that a unit loss of goal realisation for one agent is a unit gain for the other. When the risk averse agent has a high degree of goal realisation, transferring ten cents to the risk neutral agent will produce an aggregate gain in degree of goal realisation, because the risk neutral agent will value the gain more than the risk averse agent will disvalue the loss.

When this is plotted out geometrically, the effect is that, when either agent becomes more risk averse, the possibility frontier becomes a convex curve – it bulges 'north

[66] Hargreaves Heap and Varoufakis, 1995: 128.

east', away from the origin.[67] To further illustrate, consider the extreme case, when Eve becomes infinitely risk averse. An infinitely risk averse agent is indifferent over all shares of the dollar she could receive which are greater than some arbitrarily small ε. If Eve's top preference is to get $\varepsilon \approx 0$ cents, and if we suppose that Adam remains risk neutral (and so his top preference is to get \$1), then the claim point is essentially on the possibility frontier – both agents can be perfectly satisfied by Eve getting (just more than) nothing and Adam getting (just less than) a dollar.[68] Risk averse agents such as Eve are more efficient at converting resources (money) to degrees of goal realisation, because they place a higher value on having low levels of resources than risk neutral agents do. It is this efficiency that leads a bargaining solution based on the monotonicity axiom to reward risk aversion. If being risk averse means that one can be more satisfied with less money, then of course there is a sense in which it is a strategic asset to become less willing to take risks.

The key to understanding Binmore's mistake is to realise that, when Eve becomes risk averse in the divide the dollar game, *Adam's position also improves*. The Kalai-Smorodinsky solution rewards *both* agents when *either* becomes risk averse, because when the possibility frontier becomes convex there is more aggregate capacity available to increase degrees of goal realisation which can be distributed between the agents.[69] However, clearly the solution can't reward both players in *monetary* terms. It can't give both Adam and Eve more money, because we only have a fixed sum, a dollar, to distribute. Binmore fails to notice what the Kalai-Smorodinsky solution allocates to *Adam* when Eve becomes risk averse, and so he fails to appreciate that the improvement which results from risk aversion is measured in terms of degrees of goal realisation, not money. Since both agents can simultaneously benefit in terms of their degrees of goal realisation, there is no sense in which the Kalai-Smorodinsky solution implies that Eve becoming risk averse advantages her *relative to Adam*. Without such an implication, Binmore's intuitive objection to Kalai-Smorodinsky evaporates.

---

[67] *Ibid*.: 119-120.
[68] See n24. This approximates the case where $K_{Adam} + K_{Eve} = 2$.
[69] Hargreaves Heap and Varoufakis, 1995: 119-120, 128.

Neither of the intuitive arguments against Kalai-Smorodinsky and monotonicity succeed. The positive intuitive argument rests on an inappropriate conflation of standards for individual and social choice. The negative intuitive argument rests on a mistake. The cinema trip example gives us an intuitive argument in favour of Kalai-Smorodinsky and monotonicity. That argument must be balanced against the persuasive, albeit incomplete, case in favour of Nash and independence offered by the procedural arguments.[70]

§6.6 A bargaining solution for the contractarian model

The purpose of this chapter has been to identify a set of tools we can use to make prescriptions to rational agents in a contractarian choice situation. I began by setting out the bargaining problem. The bargaining problem arises when there is the possibility of productive coordination between agents, but there are various patterns that coordination could take, and the agents differ in their rankings of those patterns. A solution to the bargaining problem is an argument showing that each agent has greatest prudential reason to coordinate on the same pattern.

I have examined three candidate solutions to the bargaining problem. I set aside the egalitarian solution because it assumes agents have a shared scale of value, something which the agents in the contractarian choice situation lack. Of the prominent candidate solutions in the literature, this leaves us with the Nash and Kalai-Smorodinsky solutions. Plausible arguments can be given for each.

In my discussion of the bargaining problem and its solutions, I've used a number of examples to supplement my argument: Max and Liz bargaining over coffee; McGee and Grasp over loans; Sid and Tony over films. One might worry that, while a bargaining solution is a sensible idea in the context of an established property regime

[70] Once again, see Vanderschraaf on the limitations of the procedural arguments (2019: 166-167; n38, *supra*).

to enforce the sale of coffee or the terms of a loan, contractarian agents lack such background institutions. Alternatively, one might worry that the examples have relied on agents who implicitly care about one another's well-being – as Sid and Tony do – but that no such fellow-feeling can reasonably be ascribed to (most) contractarian agents. If the application of a bargaining solution relies on either of these conditions, the solution will not be available to the pre-political, atomic agents the contractarian model employs.

These worries are misplaced. While the examples tell stories set in benevolent or political contexts, the application of a bargaining solution doesn't require such a context. The reason Liz hands over her money in exchange for Max's coffee is not that she is afraid she will be punished by the state if she simply takes the coffee from him, or that she is independently committed to respecting Max's rights or promoting his well-being. Rather, by hypothesis, Liz lacks the power to take the coffee by force, just as Max lacks the power to take Liz's money. As discussed in §5.5.1, a core assumption of the contractarian model is that the agents are in the circumstances of justice with one another – they lack the power to simply impose their preferences on the world (including their preferences about the good or justice). The exchange of money for coffee is a means by which each agent goes about maximising the degree to which his or her goals are realised, given these circumstances.

To see that a bargaining solution makes sense, even in a pre-political choice situation with antagonistic agents, we can strip out the context built into the earlier examples.[71] Suppose that Gus and Hector are the only two agents in the world, and that they hate one another. They are engaged in a violent struggle for control of the world's resources, and spend their time raiding one another's holdings or preparing themselves against the other's raids. We need not assume that they are equally powerful, but I will suppose that neither is sufficiently powerful to destroy the other outright.[72] Over time, a stalemate develops, an equilibrium in which Gus and Hector will each expend such efforts and resources on predation designed to coercively

[71] See Buchanan, 1975: 17-34.
[72] For further discussion of this assumption, see §5.5.2.

acquire resources claimed by the other, and on defence against the predation of the other, as is profitable, up to a limit dictated by the marginal return on investment in such activities.

Gus and Hector's position is suboptimal. Each is investing resources in predation and defence which could otherwise be consumed. There thus exists the possibility of productive coordination between them. If they agree to refrain from attacking one another, they could both benefit. The 'natural', no-holds-barred equilibrium specifies a starting point for any further agreement – it is the baseline against which prospective bargains will be measured. Given this point (S_A), and a range of possible ways of distributing the benefits of coordination ($\{S^i_B\}$), a bargaining solution will select one pattern of coordination (S^*_B) as the pattern each of Gus and Hector has greatest prudential reason to agree to realise. Thus, we can see that the application of a bargaining solution requires neither goodwill on the part of the agents, nor an existing regime of rights and duties.

Chapter 7

My discussion of bargaining theory has been quite general. I now want to return to the more specific problem with which I began: Suppose we are living in a world characterised by (among other things) pluralism, moderate scarcity, and the absence of coordination. We are attempting to pursue conflicting ends, but there is the possibility of productive coordination between us. Thus, there is a bargain to be struck between us. Given the preferences people typically have, we will probably be looking for a bargain which includes a number of key features: rules governing the use of force, a property regime, a system of enforcement. If we were ideally rational, what are the terms of the bargain we would strike?

In order to make this problem more tractable, I abstract from many of the complexities which could arise at this point. Rather than attempting to capture the content of each citizen's ends, I simply suppose that each citizen has some ends. Because the model I am using is prescriptive, setting out to tell us what a rational agent would do in our place, it requires that citizens' various ends are ordered according to certain principles. When citizens' ends are so ordered, we can construct a cardinal measure of the degree to which their goals are realised under any particular patter of coordination which is available to them – a von Neumann-Morgenstern utility function. Thus, for any possible pattern of coordination, we can say how successful a citizen has been at advancing her ends. The currency in which this success is measured is the citizen's degree of goal realisation. The model supposes that agents are aiming to realise their goals to the greatest degree possible.

The substance over which citizens are bargaining is, therefore, degrees of goal realisation. More specifically, the goals in question are of a particular kind – they are goals which are realised when another agent acts in the way one prefers she acts, when she could have acted otherwise. We can call this kind of goal realisation

'political', and preferences over the goals thus realised *political preferences*.[1]

Political preferences can be distinguished from two other kinds of preference. First, consider my preference that it not rain this afternoon, rather than that it rain.[2] This is an eventuality over which (I assume) no other agent has control. I will be unhappy if it rains this afternoon. There is a sense in which my goals will be realised to a lesser degree if it rains this afternoon. But neither of these is relevant to the kind of goal realisation at stake in a contractarian bargaining situation. Whether or not it rains is a brute fact about the world – it is not something that you and I can determine through coordinating (or failing to coordinate) our actions; it is not political. It is not, therefore, part of the surplus over which we can bargain.

Second, it may appear that material resources in the world – coconuts, coffee, and the like – can be part of the bargaining surplus in a contractarian model. But, strictly, this isn't so. When a contractarian bargain allocates certain coconuts to me, what is meant is that you and I agree to act so as to bring it about that I possess the coconuts. The content of the bargaining surplus might include a requirement that you gather coconuts and give them to me, or that you refrain from interfering when I gather coconuts, or that when I have gathered coconuts you refrain from taking them, or that you research a device for gathering coconuts, or that you contribute to a system of taxation which facilitates any of these things. While we often talk as if material resources make up the content of a bargaining surplus, this is a shorthand. A surplus strictly consists only of agents' actions, because those are the only things that any agent is in a position to determine.[3] Which is to say, the terms of a contractarian bargain consist of permissions and requirements to act in particular ways – of a system of *laws*, including a system of (legal) *rights*.[4]

---

[1] §2.1.

[2] Compare Dworkin, 1981a: 192, 201-204. Note that Dworkin uses the term 'political preferences' in a somewhat different way (*ibid*.: 197).

[3] Gaus, 2003: 78.

[4] Barry, 1995: 34. Contractarian authors are sometimes at pains to point out that rights cannot be the product of a bargain, because bargaining presupposes that the bargaining agents have rights; specifically, that they have rights to the things they offer one another as part of the bargain (Gauthier, 1986: 222; Narveson, 1991: 135). It is true that the agents have to have *control* over what they offer in bargaining. Bargaining is meaningless if the agents can't deliver. One can call this control a 'right'

What a bargaining theory does is tell us which of our available patterns of coordination – which set of laws – we each have greatest prudential reason to choose. Given that we are rational, and given that, by definition, we can't be forced to coordinate with one another, when we do agree to coordinate, we are acting on the basis of our strongest prudential reasons. Given certain other conditions, the most prominent theories of bargaining prescribe an equal division of the surplus generated by coordination. To put that another way, the contractarian model tells us that the terms of the social contract (prudentially-)ought to be such as to create and maintain institutions which advance to an equal degree those goals of each citizen with respect to which the citizen has political preferences.

This is the foundation of my argument that the contractarian model prescribes liberal political institutions. In Chapter 4, I argued for an authority-based conception of liberalism. On this view, political institutions are liberal if they are the institutions which would be chosen by suitably idealised agents – agents who are rational, free, and equal – under suitably idealised conditions. Our question is now: Are the agents who populate the contractarian choice situation rational, free, and equal in the right ways, and does the choice situation meet the conditions which liberalism requires? My claim is that the agents and circumstances described in Chapters 5 and 6 meet the required conditions, and do so without appeal to any moral premises; thus, the political institutions prescribed by the model will be liberal by the lights of an authority-based conception of liberalism. I defend this claim in §§7.1-7.

However, in Chapter 4 I also conceded that the constructivism of the authority-based conception is constrained to some degree by our conventional understanding of what it means for political institutions to be liberal: that they allow individuals personal freedom of a relatively wide scope, that they recognise that each individual is equally so entitled, and that they do not promote sectarian conceptions of the

---

if one likes. Doing so makes sense for moral contractarians, because their project is to derive morality from rational choice (see §1.5, *supra*). However, the kind of rights which I take to be the product of the contractarian bargain are simply legal rights. I make no claim that they are moral rights.

good. This, then, presents us with a second question: Do the political institutions prescribed by the contractarian model fall within a reasonable margin of institutions we would conventionally describe as liberal? I address this question in §7.8.

§7.1 Rationality, individualism, and universalism

Both the authority-based conception and the contractarian model are committed to understanding their choosing agents as rational individuals. In Chapter 5, I have described the manner in which the model defines its agents as rational. While there is room for disagreement about the nature of rationality, I take it as evident that the model meets the rationality requirement of authority-based liberalism.

The model is also consistent with liberalism's commitment to understanding agents as rational *individuals*, and with understanding the rationality of those individuals as a universal characteristic. Agents in the model are atomic individuals, motivated solely by the pursuit of their own ends (though the model allows that the agents' preferences can be tuistic).[5] The model incorporates no sense in which the agents belong to pre-political groups. The prescriptions of the model are addressed to each agent individually, and the model requires that each individual agent be shown, by her own lights, to have greatest prudential reason to coordinate on a bargaining solution. The model operates with a universalistic conception of rationality. It does not suppose that the nature of rationality varies according to one's cultural or temporal context.[6] The prescriptions of the model are generated solely by the interaction between the agents, their preferences, and their circumstances – cultural and temporal context does not shape the content of the prescription (except insofar as such things are reflected in the preferences of the agents).

[5] §5.1.
[6] Compare Moehler, 2018: 28.

§7.2 Freedom, equality, and austerity

The claim that the agents in the contractarian model are free and equal is more controversial. It is nonetheless sustainable. The agents in the model are free in the sense that they are acting on the basis of their own strongest prudential reasons – each is choosing rationally in pursuit of what she herself thinks is best. By definition, the agents in the model cannot be forced to move away from the non-agreement point, S_A. On the contractarian account, it is impossible for any two agents to exit the state of nature with respect to one another without each of them choosing to do so.[7]

Importantly, this does not mean that agents cannot be threatened or harmed in the contractarian choice situation. The case of Gus and Hector shows that they can be.[8] Thus, the sense in which the agents are free is austere. There is a worry, then, that the circumstances in which the agents in the model are bargaining are insufficiently secure or peaceful to meet the standards for free choice required by an authority-based conception of liberalism.

However, in §4.2, I argued that an austere conception of freedom will suffice. Any authority-based conception of liberalism must employ agents who are motivated to agree on a set of political institutions. For rational agents to be motivated to enter political society, they must be able to gain something by doing so that they cannot gain by remaining in the pre-political state. Given that political society comes with certain costs – constraints on one's behaviour, and requirements to contribute to enforcement mechanisms – the benefits realised by political society for each agent must be significant. If, by hypothesis, the benefits are significant and can only be realised by political society, then agents in the pre-political state are necessarily faced with a kind of threat: If they do not enter political society, they will be denied the significant benefits which result. *Any* authority-based conception of liberalism is

[7] It is worth emphasising again that, at this point, I am neglecting the possibility of partial coordination. I take up that problem in Chapter 8; see also §6.1 n10.
[8] §6.6.

thus necessarily committed to a conception of freedom that is to some extent austere. Austerity is not a particular problem for a contractarian account of liberalism.

It might be objected that, given the necessity of an austere conception of freedom, it is plausible to understand the simple choice to enter political society as free, but that the choice of particular political institutions – a particular pattern of coordination – is not free if it is shaped by threats, exploitation, or differences in power between the bargaining agents. In §4.2, I argued that the extent to which we think the possibility of making threats or exploitative offers undermines the freedom of the choice of political institutions is proportional to the degree to which the agents are not equal. Agents who are equal in the relevant respects can come to free agreements with one another, even if there is nothing to stop them threatening or exploiting one another, because they are equally well-positioned to make any such threats or offers – there is equality of 'threat advantage'.[9] Thus, even a maximally austere conception of freedom of the kind employed by the contractarian model is sufficient to satisfy the freedom requirement of authority-based liberalism if it is combined with a sufficiently robust conception of equality.

In §6.4, I identified five different respects in which agents in the contractarian model might or might not be equal. They might be equally rational, equally well-informed, equally well-endowed, might have equal prospects, and might have equal bargaining power. If all five of these conditions obtain, I take it the agents will be equal in any respect a liberal egalitarian might hope them to be. However, one might reasonably doubt that this kind of equality is an accurate representation of the world in which we actually live, and the purpose of the contractarian model is to give prescriptions to real citizens which are applicable to the actual circumstances in which they find themselves.[10]

[9] Rawls, 1999a: 116, 122.

[10] We might, of course, have moral reasons for representing agents in a constructivist choice situation as equal, even if they are not actually equal. This is the contractualist position (§1.4). However, reasons such as these are not available to a contractarian theory. Contractarianism begins from the assumption that we do not agree on even very basic moral foundations, but that we nonetheless cannot implement our preferred conceptions of justice without the assistance or tolerance of other people.

There is, therefore, a possible tension in the contractarian account of liberalism I am proposing. On the one hand, the plausibility of the model as an authority-based conception of liberalism depends on the agents in the model being free. The austere conception of freedom which the model employs is only compatible with liberalism if the agents in the model are also equal in certain substantive respects. Thus, the plausibility of the model as an authority-based conception of liberalism depends on the agents being represented in the model as equal. On the other hand, however, any departure a contractarian model makes from an accurate representation of the world must be justified as a worthwhile simplification, as a reasonable presumption in conditions where we lack full information, or as irrelevant to the features of the world the model is designed to capture.[11] A representation of the agents in the model as equal which cannot be justified under one of these three headings is incompatible with the model being genuinely contractarian. Clearly, I must take care to avoid conveniently assuming a form of equality in the model, in such a way as to meet the requirements of the authority-based conception of liberalism, if the form in question cannot be defended within the limits of a contractarian theory.

So, first, in order for the contractarian model to meet the freedom requirement of the authority-based conception of liberalism, the agents in the model must be substantively equal in certain respects – specifically, they must be equal in their threat advantage. Second, equality is itself a constitutive component of the authority-based conception in its own right. In the next four sections, I will discuss the first four of the five forms of equality which the contractarian model employs. I will argue that, insofar as the model represents the agents as equal in each of these respects, doing so is defensible on contractarian grounds.

Each of the four respects could plausibly be thought to affect the agents' threat advantages, but the most important element of threat advantage is bargaining power. In Chapter 8, I will argue that the contractarian model is committed to representing

[11] See the discussion of prescriptive models in the Interface.

the agents in the model as having equal bargaining power. In combination with the discussion below, this will establish that the agents in the contractarian model are equal in the respects required by the authority-based conception of liberalism, and hence that they are also free in the required respect.

§7.3 Equal and ideal rationality

Assuming equal (and ideal) rationality is a standard feature of contractarian accounts. As I noted earlier, the fact that we are looking for a prescriptive model means it is unproblematic to assume one agent – our protagonist, so to speak – is ideally rational.[12] The question, then, is whether or not to assume that the other agents in the model are similarly rational; and, if not, whether they differ in their degree of rationality.

There are three reasons to assume equally ideal rationality for all the agents, despite the fact that this is unlikely to be an entirely accurate representation of the real world. First, given the purpose of the model, the assumption is a sensible precaution in the absence of full information about the rationality of other agents. By assuming that all agents are ideally rational, the model assumes that each agent acts according to a strategy that is a 'best response' to the strategies of each other agent.[13] The model tells me how I can best advance my ends, assuming that no other agent is making any mistakes in attempting to advance her ends which I can exploit. This may mean that I miss opportunities to exploit the irrationality of others, but it ensures that I will not be exploited by anyone else.[14] There is, thus, a maximin aspect to assuming that one has competent adversaries. If I don't know how competent others are, it is expedient to assume that they are competent.[15]

Second, the assumption of equally ideal rationality can be justified as a first-cut

[12] See the discussion of the parameters of this idealisation in the Interface.
[13] Osborne, 2004: 36-37.
[14] Binmore, 1994: 211 n34. See the discussion of the exploitation problem in the Interface, *supra*.
[15] Compare Binmore, 1994: 21, 27.

idealisation.[16] On this view, we start with a simplified representation of the world, and develop a model to cover a general case. In a similar way, me might initially develop a model to calculate the trajectory of a golf ball which neglects the influence of the wind. Inaccuracies in the representation can be rectified subsequently by theorists who wish to focus more precisely on non-ideal details, and who have access to more specific data about the agents with whom one will be interacting. That we might later want to refine the model presented here does not undermine the value of making a first cut.

Given that we have reasons to assume all of the agents in the model to be ideally rational, the conclusion that they are equally rational follows immediately. However, even if we drop the ideal rationality assumption, we might still defend an equal rationality claim as a first-cut idealisation, on the basis that at this point we lack specific information about the ways in which the agents with whom we interact differ in their degree of rationality.

Third, and finally, it is not clear that assuming all agents to be ideally rational is a particularly inaccurate representation of the world. Certainly, human beings are not always rational, and there are some kinds of cases (calculating probabilities, for instance) where we can observe systematic failures in rationality. But there is also evidence that people often approximate rationality to a reasonably close degree. How strong an assumption one finds the equally ideal rationality premise to be will depend, in part, on how demanding a requirement one thinks ideal rationality is. The formal specifications given in §5.1 and §5.2 strike me as relatively modest.[17]

§7.4 Equal and ideal information

As with the equal rationality assumption, there are two parts to a discussion of the claim that the agents in the contractarian model are equally well-informed. First, is

[16] *Ibid*.: 18.
[17] See Moehler, 2018: 105, 168; and the discussion of idealisation in the Interface, *supra*.

it reasonable to assume that the agents have perfect information? Second, do they differ in their information? As before, if we assume perfect information, equality of information follows immediately.

It is worth pausing to consider what kind of information we're talking about here. First, there is the information discussed in, or implied by, §5.5. For the model to function, the agents need to know of one another's existence. This includes the requirement that they know of one another's agential capacities – if the model is going to function, agents can't believe themselves to be surrounded by zombies. The agents must also know that they are in the circumstances of justice with one another, and how they can go about affecting the realisation of one another's goals (at least to some extent). The most straightforward way to facilitate this is to assume they can communicate effectively with one another.

I take it these informational requirements are modest, and that it will be relatively uncontroversial to assume them.[18] Things get more complicated when we turn to the information assumed by the operation of the bargaining theory. As I have set out the theory, quite a lot of information is taken to be public. Some of this is directly observable by the agents – for instance, the agents can all see for themselves what the world looks like in the absence of any agreement to coordinate, and so have direct knowledge of S_A. But, in order to calculate S^*_B, the model assumes we know each agent's vNM utility function.[19] While each agent has direct access to her own utility function, none has direct access to the utility function of anyone else. Thus, there is the possibility that agents could attempt to secure more of the proceeds of coordination by misrepresenting their utility functions to one another.[20]

Some of the arguments from §7.3, in favour of assuming equal and ideal rationality, can be recycled at this point. First, it will often be a prudent precaution to assume that the agents with whom one is interacting are well informed. Second, assuming

[18] However, compare Hume, "Of the Original Contract" (1948: 360).
[19] Nash, 1950: 155; Nash 1953: 130; Thrasher, 2014: 696-697.
[20] *Ibid*.; Moehler, 2018: 166-167. See also the discussion of fraud in §2.1, *supra*.

perfect (and hence equal) information simplifies the model greatly; this may be justified as a first-cut idealisation.

However, some of the earlier arguments will be less persuasive here. In this case, the fact that the contractarian model is prescriptive doesn't immediately speak in favour of assuming that any agent is ideally well-informed. Even an ideal poker player will almost always be acting with incomplete information, and it makes sense for us to want a poker model which does not assume away the problem of not knowing what cards everyone else is holding. Moreover, while I maintain that ideal rationality is a less demanding standard than it is often taken to be, the same cannot be said for assuming the agents in the model have complete information about one another's vNM utility functions.[21] The latter is a demanding standard, and one that is unlikely to be satisfied.

Thus, it is costly to the model – in terms of its accuracy – to assume perfect information. Nevertheless, it doesn't follow that it is similarly costly to assume equal information. If we have good reasons to model agents as ideally rational, and if we make the modest assumption that the agents know they are not perfectly well-informed, we can ask how an ideally rational agent would respond to the fact of imperfect information, and in particular to the fact that she lacks full information about other agents' vNM utility functions. I do not propose to discuss strategies for responding to imperfect information in any detail here; my point is simply to note that any such strategy will be equally available to any ideally rational agent. Moreover, while it *is* plausible to think that human beings will vary in the degree to which their utility functions are 'transparent' or 'opaque' to other people, it's less clear that, over a significant amount of time, very many people will be able to conceal the rough dimensions of their preferences. Over time, observed behaviour will tend to reveal preferences.[22] Both the fact that rational agents can employ strategies to account for imperfect information, and the fact that information about

[21] Nash, 1953: 130.
[22] See, however, Moehler (2018: 58, 171; compare *ibid*: 166-167), who notes that this is less likely to be true in 'one-shot' interactions, which might be more common in large and diverse groups of agents.

preferences is difficult to conceal over time, will be likely to constrain the degree of inequality which results from differences in the capacities of agents to misrepresent their preferences.

There is one further consideration to note at this point. We can distinguish between two functions the contractarian model serves. The primary function is prescriptive – the model is designed to tell us what we have greatest prudential reason to do when acting in the political domain. However, it also has a secondary function.[23] It provides us with a tool for evaluating our existing (or prospective) political institutions. Sometimes we'll be interested in the prescriptive function, and other times in the evaluative function. But each function provides different standards for modelling the information agents have.

From the point of view of the prescriptive function, each agent has reason to act *as if* she had perfect information. She has reason to act as if whichever of her beliefs she takes to be most reliable are true. Even when allowing that she may be mistaken in her beliefs (surely a sensible precaution), and therefore adopting a strategy which includes a margin for error, the beliefs she takes to be most reliable will anchor that strategy. We can only ever act on what we take to be the best approximation of reality which is available to us. If we're wrong, we'll end up making a mistake about what we objectively have greatest prudential reason to do; but our actions will still be subjectively justified in full.[24] Thus, from the point of view of the prescriptive function, it is reasonable for agents to model themselves as having perfect information (or good information, supplemented by a suitable contingency for error). And, once an agent assumes herself to have perfect (or good) information, it is a sensible precaution to assume that other agents are similarly well-informed, or (being rational) have taken similar precautions against not being well-informed.

From the point of view of the evaluative function, things are simpler. When evaluating a set of political institutions, what we're interested in is whether the

[23] §5.6.
[24] Compare §2.3.1.

institutions do, in fact, correspond to what the agents in the model objectively have greatest prudential reason to choose. Assuming the agents are ideally (and, hence, equally) well-informed follows immediately.

To sum up, then, the status of the equal information assumption is complicated. Assuming equal and ideal information is most straightforwardly justified as a first-cut idealisation, but it is also justified in the contractarian model's evaluative mode, because such idealisation is part of the purpose of that mode. In the prescriptive mode, as one moves away from a first-cut idealisation additional complexities will arise, and some of these will lead to inequalities between the agents in the model. I have suggested that these inequalities will be limited by the equal rationality of the agents, and by the difficulty of concealing one's vNM utility function over time. But a *fully* accurate contractarian model is unlikely to include agents with fully equal information.

Finally, it is worth noting what is *not* assumed, even when I refer to 'ideal' or 'perfect' information. As explained above, I am referring to information about the circumstances of justice and the inputs of the bargaining theory. I do not mean to suggest that the agents in the model have perfect knowledge of the future, or perfect knowledge of morality or metaethics. To take the last point, in particular: the contractarian account starts from the assumption that there is disagreement about morality and metaethics; it would be self-defeating to assume away this disagreement when making the equal (or equal and ideal) information assumption.

§7.5 Equal endowments and the contractarian baseline[25]

The foundation of the contractarian argument for liberalism is the equal division of the proceeds of coordination which is recommended by the bargaining theory under certain conditions. However, the extent to which such a division actually leads to

[25] The argument of this section draws on my MPhil thesis (2006).

an egalitarian society also depends on something else. A rational bargain guarantees each participant an improvement on the pre-agreement situation; thus, the 'holdings' of any citizen in a contractarian society will be the (equal) share of degree of goal realisation allocated by the bargaining theory, *plus* whatever degree of goal realisation the citizen was (hypothetically) able to secure in the precontractual situation.[26] Given that people differ in the degree to which their goals are realised pre-contractually, and given that the contractarian procedure preserves any such inequalities, the fact that the proceeds of coordination are divided equally does not entail that the contractarian argument leads to an egalitarian distribution overall. On the contrary, contractarian theories that have actually been proposed have tended to have a libertarian inflection.[27]

The extent to which contractarianism prescribes an inegalitarian distribution of degree of goal realisation will depend on the ratio of the degree of goal realisation any agent can secure in the relevant precontractual state to the degree of goal realisation she derives from membership of political society. The smaller the degree of goal realisation, for any agent, that can be secured in the precontractual state, relative to the degree secured by political society, the more egalitarian the overall distribution. At the limit, the contractarian procedure approaches a perfectly egalitarian distribution as precontractual degree of goal realisation approaches zero for all agents.

Thus, the degree to which contractarianism prescribes egalitarian outcomes depends on the degree of goal realisation agents can secure in the precontractual state. This, in turn, depends on how we conceive of the precontractual state. At one extreme, we have a hyper-Hobbesian state of nature, in which we take life to be so nasty, brutish, and short that agents will enjoy only a very low degree of goal realisation (with the exception of those rare agents whose idea of a perfect world is one in which they must survive from day to day on their wits and their strength). A

[26] Gauthier, 1986: 194.
[27] See, most obviously, Narveson (1988), but also Gauthier (1986, 1997) and Buchanan (1975), albeit less explicitly. I take the contrast here to be between egalitarian liberalism and libertarian liberalism; that is, to lie *within* liberalism (Narveson, 1988: 8-12).

contractarian bargain premised on such a world would be highly egalitarian, because the gains produced by political society would dominate the effects of inequalities carried over from the precontractual state. A more benign conception of the precontractual baseline, by contrast, implies a more inegalitarian social contract. Somewhat counterintuitively, then, the worse we assume the precontractual state to be, the closer a contractarian procedure will come to prescribing a liberal egalitarian political society.

My own sense is that the benefits of social coordination are substantial. When I contemplate the degree to which my goals would be realised in a lawless failed state, or in a world where no technological advancement or capital accumulation could take place, I tend to think that the proportion of the degree to which my goals are realised that derives from social coordination is very high indeed.[28] In this respect, I imagine I am quite typical. The austere conception of freedom, in allowing agents to use force against one another in the precontractual state, makes it likely that precontractual degrees of goal realisation will be very low on my account.[29] But this has the effect of making the overall, post-contractual distribution of degree of goal realisation highly egalitarian. Thus, while conceding that the contractarian procedure preserves precontractual inequalities, I maintain that political institutions constructed on the basis of the procedure would be egalitarian to a close approximation.[30]

[28] Moehler, 2018: 97-100.

[29] Note that this does not amount to making (what is sometimes taken to be) the Hobbesian assumption that people are essentially equal in their physical capacity to harm one another (see *Leviathan*, Chapter 13 (1996: 86-87)). The agents might vary widely in this regard, and it might still be the case that, for each of them, the degree to which their goals are realised in the precontractual state is a small fraction of what is available to each in political society. Compare Narveson's discussion of the distinction between 'productive' and 'predatory' powers (1991: 136-137). Strictly, we should note that Hobbes himself doesn't assume that each individual is equally capable of harming each other – the equality to which he refers (*op. cit.*) includes the possibility of agents with lesser individual capacities for harm coordinating with one another to magnify their capacities (what I have described as a case of partial coordination – see n7, *supra*).

[30] On this point see Gauthier (1974, cited in Barry, 1989: 250). It is perhaps also worth noting that something like a Rawlsian procedure would not result in perfect distributive equality either. Distributive inequality follows from the priority of the basic liberties, and the fact that the principles of justice govern social primary goods, leaving advantages that derive solely from natural primary goods where they fall (Rawls, 1999a).

The characterisation that is given of the initial bargaining situation thus plays an important role in generating the ultimate prescriptions of a contractarian model. Giving a characterisation of the initial situation is also referred to as defining a model's 'baseline', because doing so defines the basic position away from which the bargain is supposed to move the agents. The contractarian model I have described takes the absence of any coordination as its baseline (S_A), and places no restrictions on the use of force.[31]

It will be instructive to contrast this baseline with two others. The Rawlsian baseline is defined by the veil of ignorance, and by the requirement that agreement on a social contract be both unanimous and uncoerced.[32] The egalitarian character of the Rawlsian contract is a product of these conditions. The imposition of the conditions is justified by appeal to certain foundational moral principles – in particular, the requirement to treat one another as free and equal (in the robust Rawlsian sense, rather than in the austere sense I have described). Given that the basic problem which political contractarianism sets out to solve is the fact that we disagree about such foundational moral principles, but nonetheless lack the power to implement our conceptions of justice unilaterally, the Rawlsian baseline is not available to the political contractarian (even though a political contractarian might well think that the Rawlsian baseline is the morally correct one). Any other moralised baseline would be similarly unavailable, for the same reason.

There is, however, a second alternative for characterising the baseline, one which claims to be consistent with contractarianism. This understanding of the baseline is offered by Gauthier.[33] Gauthier rejects the no-holds-barred Hobbesian approach, and argues instead that a contractarian baseline is properly constrained by a 'Lockean Proviso'. The Proviso requires that, when (hypothetically) bargaining in the precontractual choice situation, agents refrain from using, or threatening to use,

[31] §6.1. Compare Buchanan, 1975: 23-26; Moehler, 2018: 101; Vanderschraaf, 2019: 120, 145.
[32] Rawls, 1999a.
[33] 1986: 190-232; 2013: 620-623. See also Danielson, 1991; Hubin and Lambeth, 1991; Narveson, 1991; Vanderschraaf, 2019: 310-320. Although Moehler criticises Gauthier's approach to the baseline (Moehler, 2018: 51-52), in certain places he seems to make quite similar moves (*ibid*.: 161-162), and his account is consequently exposed to similar objections.

force against one another, so as to ensure that the contract which emerges does not preserve advantages and disadvantages which arise from predatory precontractual behaviour.[34] This constraint is not justified by appeal to moral reasons, as in the Rawlsian case. Instead, the claim is (i) that it would be *irrational* to *comply* with a bargain reached from a Hobbesian baseline; (ii) that it would be irrational for agents to agree to a bargain if they did not believe other parties could rationally comply with it; (iii) that it would be rational to comply with a bargain reached from a Lockean baseline; and hence, (iv) that the appropriate baseline for a contractarian model is Lockean, rather than Hobbesian.

Gauthier illustrates his argument using a fable.[35] Suppose there is a society in which there are two groups – masters and slaves. The masters live lives of leisure and ease, interrupted only by the occasional need to beat their slaves. The slaves are coerced to serve the masters. One day, one of the more enlightened masters notices that his situation is suboptimal. He and his fellow masters are investing time and effort coercing the slaves. This time and effort could otherwise be devoted to more pleasant pursuits. Meanwhile, in addition to the burdens of their labours, the slaves are suffering through being beaten and kept in chains. There is thus the possibility of productive coordination between the two groups. If the slaves were to agree to serve the masters voluntarily, there would be no need to beat them or to chain them. The masters would benefit by having more time and resources for leisure; the slaves would benefit from no longer being coerced (and, indeed, could be given some further share of the proceeds of coordination, such that their material position improves). The slaves would still be much worse off than the masters, of course, but both groups would be better off than under the *status quo*.

Gauthier's enlightened master takes his proposal to his fellows and wins their approval. The masters then put the proposed bargain to the slaves, who enthusiastically endorse it. Does the society settle down happily to an even easier

[34] More precisely, the Proviso requires that an agent not advantage herself through disadvantaging another agent, so it has broader application than I suggest here (Gauthier, 1986: 206).
[35] *Ibid*.: 190-199, 227-232.

life for the (former) masters, being voluntarily served by their (former) slaves? Unsurprisingly, no. Once their masters' apparatus of coercion is dismantled, the former slaves reconsider their position. Now that the former masters lack the power to coerce them, voluntary servitude amounts to a transfer to the former masters, from the former slaves, for which the latter will get nothing in return. Given that the former slaves aren't getting anything in return, they have no reason to make such transfers. It is therefore irrational for them to do so. Thus, it is irrational for them to comply with the bargain.

We need to be careful that we understand Gauthier's argument correctly here. His claim is not simply that, once the masters have done their part, the slaves lack a reason to do theirs. The whole point of his theory of compliance is to solve problems of diachronic contribution: If we both need one another's help harvesting our crops, but your crop is ripe now, and mine will not be ripe for a month, we have a problem of diachronic contribution. We would both benefit if we could agree to help one another in turn, but once we have harvested your crop you will lack any reason to return the favour. I know this, and refuse to help you, and so you refuse to help me.[36]

Gauthier's solution is to say that it is (possible and) rational for us to adopt a *disposition* – constrained maximisation – which gives us an exclusionary reason to comply with rational bargains we have made, so long as others similarly comply, even when we lack non-dispositional reasons to comply.[37] In the harvest case, if we are both constrained maximisers (and both know one another to be constrained maximisers, etc.), we can benefit by agreeing to cooperate. I know you won't cheat me, because when the time comes I know you will have an exclusionary reason, based on your disposition, to help me harvest my crops if I have helped you harvest yours. Therefore, I have a reason to help you now, and because I do so, you have a reason to help me later. Having the disposition of a constrained maximiser leaves us both better off than we were when we each had the disposition of straightforward

[36] See the Introduction n35; Vanderschraaf, 2019: 3-7, 25-30.
[37] Gauthier, 1986: 165-170; see also Moehler, 2018: 51, 56-58. On exclusionary reasons, see Raz (1999: 35-48).

maximisation. Thus, it is rational to be a constrained maximiser rather than a straightforward maximiser when it comes to first-order compliance with bargains. And this is justified by the standards of second-order straightforward maximisation – we actually maximise most effectively by *not* applying straightforward maximisation to first-order situations.

So, according to Gauthier, if the masters and the slaves are rational, they will be constrained maximisers; and if they are constrained maximisers, the mere fact of a diachronic contribution problem won't faze them. Gauthier's argument that it is irrational to comply with bargains reached from a coercive baseline therefore relies on something else. Unfortunately, it is not entirely clear what this is. But, there seem to be two plausible ways of understanding his reasoning.[38]

First, Gauthier's claim could be that it would be irrational for an agent to dispose herself to comply with bargains reached from a coercive baseline because "this would simply invite others to engage in predatory and coercive activities as a prelude to bargaining."[39] The argument here seems to rest on a worry about creating counterproductive incentives in one's prospective bargaining partners. Thus, Gauthier is claiming that it would not be rational for the slaves to adopt a disposition to comply with bargains reached from coercive baselines, because they could do better with a different disposition.

Second, Gauthier claims that it would be irrational for an agent to dispose herself to make voluntary unproductive transfers to others. An unproductive transfer is an action which "involves a utility cost for which no benefit is received, and a utility

[38] Gauthier, 1986: 193-199. I neglect a possible third line (implicit at *ibid*.: 195-196), where Gauthier wonders about the prospect of coercively-based coordination in apartheid South Africa or communist Poland. We might take him to be appealing to reflective equilibrium here – claiming that the coercive baseline cannot be correct because it gives intuitively incorrect solutions in particular concrete cases. But, rhetorical flourishes aside, this is not in keeping with the rest of Gauthier's methodological commitments. Similarly, it might appear (at *ibid*.: 197-198) that a conceptual argument about the nature of cooperation is being offered. Again, though, it is hard to see how such an argument could get very far within Gauthier's larger framework. Finally, there are hints, but no more, about a concern with the practical stability of coercively-based coordination (*ibid*.: 196, 198, 230; see also §3.2.2, *supra*). I will not pursue these possibilities here.
[39] Gauthier, 1986: 195.

gain for which no service is provided."[40] The transfer is a cost for the agent providing it. As such, it can be rational to provide it "only in so far as it is coercively extracted."[41] Here, the point is that the bargain reached from the coercive baseline involves unproductive transfers from the former slaves to the former masters, and such transfers are irrational inasmuch as the former slaves gain nothing from them. Thus, it will not be rational for the former slaves to comply with a bargain reached from the coercive baseline.

Now, Gauthier's argument at this point is clearly going to depend in some way on his theory of compliance, and I've already expressed scepticism about the prospects for that theory. But suppose, for the sake of argument, that we accept the theory behind constrained maximisation. What follows?

The upshot, Gauthier argues, is that once the coercive apparatus of the masters has been dismantled, the former slaves simply stop serving the former masters and go about their own business based on a pattern of coordination to which they and the masters would have agreed had the slaves not been coerced in the initial bargaining situation. From this, we are to conclude that the initial bargaining situation for a contractarian model must be characterised in a way that excludes the influence of coercion. Agents who are coerced in the precontractual state – the slaves – will have no reason to comply with political institutions which formalise their dispossession. It is not rational to agree to a bargain with which one's fellow participants have no reason to comply. Therefore, bargains based on a coercive precontractual state are not available to rational agents. Would-be coercers – the masters – know that it would be pointless to offer their victims terms of coordination which preserve gains brought about by coercion. Nonetheless, so long as it is still the case that productive coordination is possible on the basis of a *non*-coercive baseline, there is no reason for rational agents to remain in the precontractual state. It is rational for coercers to offer, and for the coerced to accept, terms of coordination which take as their baseline a hypothetical world in which no coercion has occurred.

[40] *Ibid.*: 197.
[41] *Ibid.*

This, then, gives us Gauthier's justification of the Lockean Proviso. The relevant baseline for a contractarian model is a precontractual state in which Lockean rights against coercion are respected. Given that agents have different productive capacities, a precontractual state in which Lockean rights are respected will be populated by agents with widely divergent degrees of goal realisation.[42] These inequalities will then be preserved by the contractarian procedure, and contractarianism will prescribe some variety of libertarian distribution.

The key move in the masters-slaves argument is evidently the claim that, once the apparatus of coercion has been dismantled, the former slaves have no reason to comply with the bargain they had made. As we've seen, this is not simply a problem of diachronic contribution. Rather, the former slaves lack a reason to comply either because it would be irrational to adopt a disposition for compliance which has the effect of incentivising one's own victimisation in the precontractual state; or, because it would be irrational to adopt a disposition to comply with bargains that include unproductive transfers.

Gauthier's argument is attractive on its face, in that it appears to allow us to rule out coercion in the precontractual state without appealing to controversial moral premises. If his argument were successful, contractarianism would not be restricted to a conception of freedom as austere as the one I have proposed. It would, consequently, have a more libertarian character, but that would only be a disadvantage to the extent one finds libertarianism morally unpersuasive.

Nevertheless, we should reject Gauthier's account of the contractarian baseline. Suppose we contrast Gauthier's fable of the masters and the slaves with another fable.[43] Suppose there is a wealthy society of gatherers. The gatherers earn their keep by picking fruit that grows from the branches of tall trees. Gathering the fruit involves climbing the trees, a feat which requires hard work and considerable talent. Any fruit

[42] Narveson, 1991: 136-137.
[43] Compare Moehler, 2018: 169-170.

that isn't gathered is eaten by birds before it can fall to the ground.

Intermingled with the gatherers is a race of beings who are like the gatherers in almost all respects. However, these beings suffer from a congenital inability to climb trees. Lacking any other source of food, they subsist on whatever fruit they can raid from stockpiles laid up by the gatherers. This isn't much, as the gatherers invest some portion of their significant wealth in guarding their stockpiles, and can do so effectively. As such, the raiders live miserable lives – insecure, impoverished, and malnourished.

One day, the raiders send a delegation to treat with the gatherers. They propose to cease their raids on the stockpiles, if the gatherers will voluntarily transfer to them some small proportion of the fruit they gather. Both parties will benefit, the delegation notes. The raiders will have a slightly higher, and more secure, standard of living, though still nothing approaching the standard of living of the gatherers. Meanwhile, the gatherers will be significantly better off; all of the resources they presently invest in guarding their stockpiles can instead be invested in exotic fruit cocktails. Being rational, the gatherers agree to the raiders' proposal.

What happens next? Suppose we follow Gauthier, and assume there is some means by which the raiders can permanently alienate their capacity to raid (perhaps they agree to have little bells attached to their necks; or GPS tracking devices implanted in their ears). With the threat of raiding eliminated, do the gatherers have any reason to comply with the bargain? On Gauthier's account, I take it they don't. The baseline for the bargain is coercive. The transfers to which they have agreed are unproductive, and adopting a disposition to comply with bargains based on a coercive baseline would incentivise the raiders to raid to the greatest extent possible in the precontractual state, so as to maximise their bargaining position *vis-à-vis* the gatherers.

On Gauthier's account, I take it the gatherers would repudiate the bargain as soon as the threat of raiding has been eliminated. They would refuse to make the agreed

transfers, and the raiders, being unable to raid, would all starve. Anticipating this, the raiders would never enter into the bargain in the first place, and the precontractual state would continue indefinitely. But is this rational? There *is* the possibility of mutually advantageous coordination between the two groups. The precontractual state is suboptimal. How can it be rational for the groups to remain in a suboptimal state? How can it be rational for either of them to adopt a disposition which would prevent them engaging in mutually advantageous coordination?

Now, I take it there is a plausibility to Gauthier's masters and slaves fable. It strikes us as morally reprehensible that advantages won through violating a person's rights in the precontractual state might be enshrined as legal privileges in political society by the logic of rational bargaining.[44] But our moral misgivings have nothing to do with the rationality of the bargaining. The gatherers and raiders fable is structurally identical to the masters and slaves fable. I have made the predatory party – the raiders – more sympathetic than the masters, by making them desperately poor relative to their victims. But that shouldn't have any effect on the logic of the argument. If we can't believe it would be rational for the gatherers and the raiders to remain in the suboptimal state, we cannot accept the logic of Gauthier's masters and slaves fable. As such, we have no (non-moral) reason to accept the Lockean baseline.

Gauthier might take issue with my claim that the two fables have the same structure. When the slaves defect from their agreement with the masters, the society does not immediately return to the precontractual state. Rather, unless the masters also defect, the society simply moves to a different pattern of coordination – a different S^*_B – one based on a non-coercive baseline. Both former masters and former slaves are still better off than in the precontractual state, it is just that the former slaves are taking a larger share of the proceeds of coordination, because the former masters are not benefiting from their precontractual predation. By contrast, in the gatherers and raiders fable, there is no possibility of productive coordination in the absence of a

[44] If you think the raiders are not violating anyone's rights in taking what they need to survive from the stockpiles, suppose they can only gain access to the fruit by assaulting and injuring the gatherers.

coercive baseline – the raiders cannot bring anything to the table other than their agreement to refrain from raiding.

This is certainly a difference between the cases, but it is not a relevant one. First, Gauthier's bargaining principles are going to have to be applicable in cases like the gatherers-raiders scenario, as well as in cases like the masters-slaves scenario. And, when applied to the gatherers and the raiders, Gauthier's principles are not rational.

Second, the principles don't work in the case of the masters and the slaves either. Suppose we imagine that the former masters have the power to reimpose coercion when the former slaves defect.[45] Gauthier considers the possibility that they would then threaten the former slaves: Comply with the bargain you made while we were coercing you, or we will enslave you again. If the former masters can make threats of this kind, then we can say that the former slaves *are* getting something in return for their 'voluntary' transfer – they are getting the assurance that they will not be returned to their chains.

Nevertheless, Gauthier tells us that any such threat would be idle. On the assumption that the coordinated outcome based on the (hypothetical) non-coercive baseline leaves both groups better off than they are at the coercive baseline, the former masters would be threatening to move themselves from a world they prefer to a world they disprefer. The former slaves know the former masters are rational, and a rational agent will not bring about a dispreferred state when a preferred state is available. Thus, even if the former masters could reimpose coercion, that possibility still wouldn't give the former slaves reason to comply with a bargain based on the coercive baseline.

But this last move by Gauthier is surely too quick. It is not the case that the former slaves can simply select any coordination point they please on the Pareto frontier, safe in the knowledge that the former masters will not revert to the precontractual

[45] Narveson, 1991: 140.

state. The bargaining theory is designed to select a rationally acceptable compromise between the masters and the slaves, based ultimately (in Gauthier's account) on the contention that in bargaining situations the party with more to lose by deadlock than by concession must rationally concede to other parties. It is not at all clear that the former masters have more to lose by a return to the precontractual state than the former slaves do, and there is certainly no reason to think it would be generally true that coercing parties have more to lose than coerced parties. In fact, the application of Gauthier's bargaining theory selected the exploitative bargain – the bargain which preserves the gains of predation – as the point of agreement when the alternative was the coercive precontractual state. Once again, then, we don't have any (non-moral) reason to adopt a Lockean baseline.

The appropriate baseline for a contractarian model is therefore the Hobbesian one with which I have been working.[46] The Rawlsian alternative is ruled out by the fact of pluralism; the Lockean alternative is ultimately unsustainable on non-moral grounds.[47] This means that contractarianism is necessarily committed to a maximally austere conception of freedom. However, the fact it is committed to the Hobbesian baseline also means that a greater proportion of the degree to which goals are realised will be the result of political coordination than under alternative baselines. When bargaining power is equal, this in turn means that the contractarian model will tend towards egalitarianism.

[46] See also §6.3 n23 and §6.4 n38: *Given* a Hobbesian baseline, we might think that the process of bargaining must itself be the object of an agreement between the agents. Combined with the assumption that bargaining is costless (§5.5.2), this may present a problem for contractarian theories employing a Nash bargaining solution (say; see the introductory remarks in Chapter 8, particularly n4). Still, my discussion here leaves open the possibility that a particular structure for bargaining is fully determined by the circumstances of the state of nature (which would mean the process of bargaining *need* not be the object of an agreement between the agents).

[47] See, however, Van Donselaar's argument that, under certain conditions, offers of rewards are more cost effective than threats of punishment (2013). I take it this argument is consistent with mine. That the Hobbesian baseline permits coercion does not imply that it requires coercion; agents will choose whatever means are most conducive to their ends.

§7.6 Equal prospects

An agent's 'prospects' are the range of possible patterns of coordination available, between her and all other agents, which are superior to the pre-agreement state (S_A) from the point of view of each party. Recall that a pattern of coordination between any two agents is an agreement between the agents to decide to act in particular ways. Each agent possesses a ranked ordering of all the possible patterns available. Recall also that, for the purposes of the contractarian model, only a subset of an agent's preferences – her political preferences – are relevant to assessing her prospects.[48]

Formally, agents in the contractarian model have equal prospects if the solution space, $\{S^i_B\}$, is symmetric. A solution space is symmetric if and only if, for any point (l, m) in the space, the point (m, l) is also in the space.[49] What this means is that, if there is a pattern of coordination available such that the degree to which the goals governed by *L*'s political preferences are realised is 80%, while the degree to which *M*'s are realised is 40%, there must also be available a pattern of coordination under which the reverse is the case.[50]

Under what kinds of circumstances might agents in the contractarian model *not* have equal prospects? We can set aside one misunderstanding quickly. Suppose Liz has a much sunnier disposition than Max. She is capable of achieving much higher levels of utility than he is. Under her most preferred pattern of coordination, her utility level would be 100 utiles. Under Max's most preferred pattern, his utility level would be 50 utiles. In this case, it might be thought that Liz and Max have unequal prospects. No matter what anyone does, Max can never achieve as high a level of utility as Liz. The misunderstanding here is in thinking that the contractarian model takes utiles to

[48] See the introductory remarks in this chapter.
[49] This can be extended to accommodate $n > 2$ agents by requiring that if $S^i_B = (i_1, i_2, i_3 \ldots i_n)$ is in the solution space, any permutation of S^i_B is also in the space. See §6.1 n12.
[50] Note that the degrees to which the agents' goals are realised in this example do not sum to 100%, because there is no need to assume the agents' goals are strictly opposed (see §6.3 n24; and §6.5 n68). Note also that the percentages used here express degrees along an individual's vNM utility function, rather than anything else (see §5.1, §5.2).

be its distribuendum. In fact, the model is predicated on it being impossible or inappropriate to make interpersonal comparisons of utility.[51] Instead, by using a vNM utility function, the model defines Liz's and Max's least-preferred and most-preferred options such that they occupy the '0' and '100' points on a scale. Liz reaches exactly the same degree of goal realisation under her most-preferred pattern of coordination as Max does under his most-preferred pattern. Thus, insofar as all points on their vNM utility functions are similarly accessible, the agents in the model have equal prospects by definition.

Is there reason to think all points on any agent's vNM utility function will be accessible? It is tempting to think of degrees of goal realisation as a distributable resource, like manna, which can simply be taken from one agent and given to another.[52] But we should be wary of such simplification. The realisation of goals can arise from distributions of material resources, but it doesn't follow from the fact that we have a perfectly divisible, costlessly transferrable resource to distribute that we can similarly achieve any pattern of degrees of goal realisation we want.[53] The question is, how do distributions of degrees of goal realisation relate to distributions of (among other things) resources? They might not have a simple, linear relationship, such that having twice as much of a resource means that one's degree of goal realisation increases by a factor of two.

Consider two possible ways in which distributing degrees of goal realisation can be complicated. First, agents might have *lumpy* preferences.[54] Suppose I am in favour of the prohibition of alcohol, and that I prefer a world in which alcohol is prohibited to any world in which it is not. I have lumpy preferences if it doesn't follow that I prefer worlds in which there are greater restrictions on the availability of alcohol to worlds in which there are fewer restrictions. Suppose, instead, that I am indifferent over all worlds other than a world in which alcohol is prohibited entirely. Lumpy preferences are preferences that have an all-or-nothing character, rather than varying

[51] §6.2.
[52] Ackerman, 1980.
[53] Compare §6.4 n30.
[54] Compare Moehler, 2018: 167.

smoothly along a continuum. Second, agents might have vNM utility functions that *covary*.[55] Suppose that, for some reason, I have preferences over the degree to which you realise your goals, and thus my degree of goal realisation is linked to your degree of goal realisation such that whenever your goals are realised to a degree less than 50%, my goals are realised to a degree which is twice the degree to which yours are realised. As we give additional resources to you, your goals are realised to a greater degree, and so too are mine, even though my share of resources hasn't changed.

In either of these cases, it will not be straightforward to realise certain points on some agents' vNM utility functions. Take the case of lumpy preferences, and suppose that I am in favour of prohibiting alcohol and that you are opposed. Suppose that, if I get my way, my goals will be realised to some high degree, and that if you get your way, my goals will be realised to some low degree. Our bargaining theory tells us to divide the available degrees of goal realisation equally, but given the structure of my preferences that might not be possible. It might be the case that I will necessarily end up with a lower or a higher degree of goal realisation than you do – the mid-range of my vNM utility function might appear to be inaccessible. As a result of my lumpy preferences, it might be the case that the point (25, 50) is available, but that the point (50, 25) is not, and thus the solution space would be asymmetric.[56]

There are two ways in which to respond to this kind of problem. The first is to appeal to the existence of alternative ways in which goals can be realised. The idea is that, as well as caring about the (un)availability of alcohol, we will also care about other things.[57] If alcohol is prohibited, you can be given a greater share of these other things to compensate you (in goal realisation terms) for the prohibition of alcohol. Alternatively, I can be given a greater share to compensate me for the availability of alcohol. The existence of lumpy preferences does not make it impossible to distribute degrees of goal realisation equally if compensation of this kind is feasible. Whether or not it is feasible is contingent on the preferences of the real agents in

[55] Compare Dworkin, 1981a: 197-198.
[56] The format here, and in the text below, is (me, you). The numbers represent degrees on a vNM utility function ranging from 0 to 100.
[57] *Ibid*.: 198; Moehler, 2018: 167.

question. It strikes me as plausible to say that most human beings will have preferences which are not strictly absolutist in the way that the prohibition example imagines, and thus compensation will probably be possible. But, as always, the contractarian theory is hostage to the underlying preferences of the people with whom we are concerned.

The second response is to shift the terms of the distribution to expected, rather than actual, degree of goal realisation. An indivisible good, such as the authority to determine the law with respect to the availability of alcohol, can be rendered divisible in expected value terms by means of a lottery. Rather than either one of us being given the authority to determine the law with respect to alcohol, each of us is given a lottery ticket. The prize of the lottery is the authority in question, and the tickets are weighted in such a way that the inclusion of my ticket in my bundle of resources, and the inclusion of your ticket in yours, renders the overall distribution of degrees of goal realisation equal.[58]

In principle, any point on an agent's vNM utility function which was initially inaccessible as a result of lumpy preferences can be rendered accessible by means of an appropriate lottery. One might object, at this point, claiming that the problem of lumpy preferences can simply be reintroduced at the level of expected value. However, such a claim would rely on me being indifferent between, for example, being given something I value with a probability of 0.4, and being given it with a probability of 0.6. And that looks irrational – if I value *X* then, leaving aside my preferences about risk itself, I must prefer achieving *X* with a higher rather than a lower probability.[59]

So, lumpy preferences don't seem to pose an insuperable problem. What about covariance? In the case described above, for a given range of your vNM utility function, my goals are realised to a degree twice the degree to which yours are

[58] Compare Thoma, 2015: 3339. We can argue, on this basis, for rights of democratic participation (equal access to political authority), perhaps constrained by other constitutional rights (limits on democratic authority which apply when high-priority interests are at stake). See §7.8, *infra*.
[59] Rawls, 2005: 83.

realised. The problem here is not that any particular point on my function or yours is inaccessible, but rather that because my function is not independent of yours, certain points on our functions are not *simultaneously* accessible. Thus, the point (35, 60) is accessible, but the point (60, 35) is not (because when your goals are realised to a degree 35, mine are necessarily realised to a degree 70, but the reverse is not true). Again, this would mean the solution space is asymmetric, and that agents can therefore have unequal prospects.

As before, it is an open question whether the real agents with whom we are concerned will actually have vNM utility functions which covary in such a way. It certainly seems *possible*, which is why I was unwilling to conclude that the Nash and Kalai-Smorodinsky bargaining solutions would be extensionally equivalent.[60] The present issue is whether the possibility of unequal prospects undermines my claim that the contractarian model represents its agents as free and equal in the manner required by an authority-based conception of liberalism. In my view, it doesn't.

The problem of covariance can be understood as a feasibility constraint on the contractarian model. The model prescribes an equal division of the proceeds of coordination (where bargaining power is equal). If, due to covariance, an equal division is not accessible, the model will prescribe the closest accessible point in the solution space. This is a compromise in the model's egalitarianism, but it is a compromise with reality. So long as we think that liberalism is compatible with a commitment to ought-implies-can, this kind of feasibility constraint does not undermine the claim that the model is a plausible example of an authority-based conception of liberalism.

One might, of course, object to the account of equal prospects I have given here. I'll consider one particularly troubling objection. In the course of discussing lumpy preferences, and how a contractarian theory might accommodate them, I raised the

[60] §6.4.

possibility of compensating an agent when the goal associated with her lumpy preference is not realised. I suggested that we could do this by transferring to her control of additional material resources (say), in order to ensure that the overall distribution of degrees of goal realisation was equal. The worry is that this procedure places no limits on the kinds of preferences for which people could be entitled to compensation.[61] Why, for instance, do crazy prohibitionists deserve compensation from the state in return for agreeing to refrain from violently assaulting people who produce, trade, or consume wine? People, that is, who are not harming anyone else; who are simply exercising their rights to profit from the creation or consumption of alcohol. We can create progressively more extreme cases illustrating the objection – why is it that Nazis are owed compensation for declining to exterminate their Jewish neighbours?

Ultimately, I will bite the bullet here. But there are some things that can be said to make the conclusion more palatable. First, if the objection is that we (morally-)ought not to compensate people when they are forced to forego preferences for violating rather than respecting the rights of others, then this case is simply a version of the Gus-Hector scenario, or the masters-slaves scenario, or the gatherers-raiders scenario.[62] The starting point for a contractarian theory is the claim that we do not agree on a conception of justice, and lack the capacity to impose our preferred conception. Thus, we do not agree on what (moral) rights people have. The bargain itself determines the (legal) rights people have in political society, and the appropriate baseline for that bargaining is a Hobbesian one in which violations of (moral) rights are possible.[63] Suppose that, instead of prohibitionists and oenophiles, or Nazis and Jews, the case is one in which a group of liberals is threatening to violently assault libertarians who are engaging in free exchange with one another and don't see any reason to pay tax. The liberals, no less than the prohibitionists and Nazis, believe they have the correct understanding of justice, and that the libertarians (no less than the oenophiles and the Jews) are acting unjustly. The

[61] Except insofar as the preferences must be political. Compare Nozick, 1974: 272-273; Dworkin, 1981a: 198; Raz, 1986: 115-116. I thank Kieran Oberman for pressing this point.
[62] §6.6, §7.5.
[63] §7.5.

contractarian theory doesn't take sides in these disputes – it asks what a rational compromise between the relevant parties would be, not which party is morally correct. It is compatible with any of the canvassed positions being morally correct.

Second, perhaps the objection is that some people – the Nazis, for example – are so extreme that no compromise with them is possible. So, even if we did not have moral objections to compensating Nazis for refraining from violating others' rights, if the Nazis' highest preference is the extermination of the Jews, there is no amount of compensation the Jews can offer the Nazis which would make up for their continued existence. But, if this is so, the scenario is one I have already discussed – it is a case of pathological preferences.[64] The contractarian theory is not committed to any compromise with agents who have pathological preferences (precisely because, by definition, no such compromise is rationally possible). In such an event, the contractarian says the same thing a Rawlsian says: You don't offer the Nazis (or prohibitionists, or liberals) reasons. You shoot them.[65]

Third, perhaps the objection is that the prohibitionists or the Nazis are getting an *unfair share* of the available degrees of goal realisation under the proposed model.[66] To illustrate, consider a variation on the Gus-Hector example. Suppose Gus is a liberal egalitarian. He has a preference (a moral commitment) for distributing resources equally rather than unequally.[67] By contrast, suppose Hector is a psychopath. He cares nothing for Gus, he simply wants as great a share of resources as possible. Thus, Gus's most-preferred pattern of coordination is one which results in resources being divided (50, 50). Hector's most-preferred pattern is one which yields (0, 100). One might think that the contractarian model would select a rational compromise between the agents at (25, 75) – that is, halfway between their ideal points. It would seem, then, that Gus is penalised by the theory for having what we think is the morally correct conception of distributive justice, while Hector is advantaged in being thoroughly amoral.

---

[64] §5.5.2.
[65] Dreben, 2003: 329; Rawls, 2005: 64 n19.
[66] Compare Dworkin, 1981a: 198-200.
[67] On the relationship between preferences and moral commitments, see §5.1, particularly n9.

There are two possible responses to this version of the objection. The first is to note that a distribution of resources at (25, 75) is not the same as a distribution of degrees of goal realisation at (25, 75). The model cares about distributing degrees of goal realisation, and it is perfectly possible that distributions of resources will not match distributions of degrees of goal realisation. Insofar as we think this is a problem, because we are morally committed to something like equality of resources as a theory of distributive justice, the appropriate response is that the contractarian theory starts from the assumption that we disagree about the correct conception of distributive justice, and lack the power to simply impose our preferred conception. Given this assumption, (25, 75) is the best distribution feasible.[68]

I take it this response rebuts the objection. Nevertheless, it does so at a cost, since it concedes that in pursuing equality of goal realisation (for goals governed by political preferences), the model may diverge from our pre-theoretic understanding of liberal distributive justice. There is, however, a second response available. Recall that the substance over which agents in the contractarian model are bargaining is control of one another's decisions to act. The apparent problem in the (revised) Gus-Hector case is that Gus appears to want to control half of the available surplus, and to want Hector to control the other half. Meanwhile, Hector wants control over the whole surplus. But this description of the situation is misleading. In being an egalitarian – in wanting Hector to control half of the available surplus – Gus is not expressing indifference as to how the remaining half of the surplus is allocated. Gus actually has preferences which cover the whole of the available surplus; there is no part of it he is uninterested in determining. Expressed in terms of control of agents' decisions to act, Gus's ideal distribution, as an egalitarian, is (100, 0) – a distribution in which he has full control over how agents in the society decide to act. He would then exercise this control in such a way as to bring about, for example, a distribution

[68] Note a special case of this version of the objection. Suppose Gus's conception of justice *is* equality of degree of goal realisation (for goals governed by political preferences). In this case, the model will *always* select a pattern of coordination which maximally realises Gus's goals. This would be an extreme example of the covariance problem. The appropriate response is thus given in the text, *supra*: covariance is a feasibility constraint on the contractarian model.

consistent with equality of resources. Properly understood, then, an egalitarian like Gus is claiming just as great a proportion of the available surplus as a psychopath like Hector is. Given that Gus is claiming (100, 0), and Hector is claiming (0, 100), the model will select a compromise at (50, 50). If each agent then acts in a self-interested way *within* the resulting political society, the distributive pattern realised will be something approaching equality of resources.[69]

To sum up, then: The possibility of covariance means that an equal distribution of degrees of goal realisation may not always be feasible. Nonetheless, agents in the contractarian model will have equal prospects in the sense that (assuming equal bargaining power) the bargaining theory will select the pattern of coordination which most closely approximates an equal division of degrees of goal realisation in cases where a strictly equal division is not available. Sometimes, this will mean compensating people for refraining from violently pursuing their conceptions of the good, or of justice. But, given the fact of pluralism, this is inevitable. Moreover, while it may appear that agents with, for example, liberal or egalitarian conceptions of justice are disadvantaged in the contractarian model relative to amoral agents, this is an artefact of the manner in which we tend to think about what it is that the bargaining process is distributing. Properly conceived, there is no such bias in the model.

§7.7 Neutrality

It remains to be shown that the contractarian model is consistent with liberalism's commitment to neutrality or toleration. I have little to add here to the argument presented in §4.2. There, I distinguished between two dimensions of liberal neutrality: justificatory and consequential. Political institutions are justificatorily neutral if and only if they can be justified to a citizen using only her own subjective reasons. The classic case of a violation of justificatory neutrality would be justifying

[69] The covariance caveats discussed above apply. See n68.

a political institution – the rule of a monarch, perhaps – on the basis that the institution has been divinely ordained according to a particular religious tradition. For any citizen who does not accept the foundational commitments of that religious tradition, the fact that the god in question (is claimed to have) appointed the monarch to supreme political authority is not a (subjective) reason to comply with the commands of the monarch. (It might nonetheless be an objective reason, if the religious tradition is correct, but justificatory neutrality trades in subjective reasons.) By contrast, consequential neutrality is the requirement that the actions of the state not have the consequence of advancing one conception of the good (held by a citizen) over other conceptions of the good (held by other citizens). An example of a violation of consequential neutrality might be the creation of a police force. In creating a police force, a state helps citizens who prioritise a life of discipline, order, and public service to pursue their conceptions of the good, but does not advance the conceptions of the good of other citizens to the same degree.

The contractarian model incorporates both conceptions of neutrality. First, the model tells each citizen that she has greatest prudential reason to converge with other citizens on that set of political institutions which is identified as the coordination point by the bargaining theory. In doing so, it appeals only to (i) the rationality of the citizen and (ii) the citizen's own preferences – the prescription of the model is justified as the most effective means to the citizen's own ends. The model does not rely on any reasons which are not subjectively available to the citizen herself. As such, the prescriptions of the model are consistent with justificatory neutrality *no matter what their content*.

Second, the contractarian model suggests that the proceeds of social coordination (prudentially-)ought to be distributed equally among the participants. This prescription relies on a number of further claims about rationality and equality which I have been defending in this chapter, and on the argument in favour of equal bargaining power which will be presented in Chapter 8. Insofar as these arguments are successful, the prescriptions of the contractarian model will be consistent with consequential neutrality. I take a citizen's conception of the good to be the totality

of her political preferences (including her preferences about the right). What the model distributes is degrees of goal realisation (for goals governed by political preferences). Thus, in prescribing political institutions which realise each citizen's goals (of those governed by political preferences) to an equal degree, the model is committed to creating political institutions which advance each citizen's conception of the good to an equal degree.

§7.8 The scope of contractarian institutions

I now turn to examining the substantive content of the political institutions which would be selected by a contractarian procedure. There is a limit to the degree of specificity with which I can discuss this question, given the level of abstraction at which I am working. One could develop a more concrete theory, filling in the specific preferences of a particular population of real people in increasing detail, and applying the contractarian procedure to yield increasingly determinate prescriptions. That task is beyond my scope. However, I can provide an outline of the institutions which plausibly follow from the procedure for populations with reasonably typical preferences.

So long as we continue to assume equal bargaining power, the guiding principle of the contractarian procedure is equality of degree of goal realisation (for goals governed by political preferences). Clearly, then, the specific political institutions prescribed by the theory will depend on the preferences of the citizens to whom it is applied.[70] If, in a particular population, it was unanimously believed that Donald Trump is the divinely anointed leader of humanity, the preferences of the citizens could well be for monarchical and theocratic political institutions in which Mr Trump exercises unchecked secular and religious authority, rather than any form of democracy. The contractarian procedure can certainly produce such outcomes, but it only does so given a citizen body with preferences which are both exotic and

[70] §4.1.

highly uniform. In contrast, the case with which I am concerned is that of a society characterised by much greater disagreement.[71]

The more central case, then, is a society in which citizens differ in their conceptions of the good, in their conceptions of justice, and consequently in their claims on material resources and state authority. In such a society, the contractarian procedure will tend to yield political institutions which are substantively liberal. We can identify a number of reasons for this.

First, rational agents interested in pursuing the realisation of their goals will be interested in securing the necessary conditions of their goals being realised. Foremost among these will typically be a certain security of their persons. Among the necessary conditions of me realising my goal of windsurfing, say, are that I am alive, that I am not being imprisoned or assaulted, that there exist people who are able to manufacture windsurfing equipment and are free to do so, etc. Thus, while the contractarian theory places no special priority on basic liberal rights and freedoms, seeing them simply as things over which agents might have various preferences, in practice many rights and freedoms will follow straightforwardly from applying a principle of equal goal realisation because basic rights and freedoms will be among the necessary conditions of realising a whole range of other goals. Note, in particular, that among the necessary conditions of me realising many of my goals is that certain other people have basic rights and freedoms.

Now, not all preferences have experiences as their object. I may, for example, prefer that my children grow up in a safe environment rather than a dangerous environment, even if I will not be around to experience them doing so. Thus, basic security – or even life – is not a necessary condition for the realisation of all goals, and some citizens may rank the realisation of some non-experiential goals above their basic security. To the extent that they do, the contractarian model may not result in citizens universally sharing standard liberal rights and freedoms. Note,

[71] §5.5.1.

however, that this will only follow in the event that a citizen's equal share of degree of goal realisation is *exhausted* by the realisation of her non-experiential goals. This would be unusual, and in such an event I do not have a strong intuition in favour of insisting on the provision of rights and freedoms for people who would (competently) prefer to exchange those rights and freedoms for other goods. Moreover, in privileging the judgement of the individuals in question, I take this to be an appropriately liberal response.[72]

Second, then, rational agents aware of the fact that their preferences may change over time will set about securing the *means* of realising their goals, as well as or instead of securing the realisation of their goals directly. I have typically discussed agents' preferences as if they were something static, but this is a simplification. While some of our preferences may be enduring, others will change over time. That is to say, citizens are likely to have plenty of *dynamic* preferences. On the assumption that it would be impractical to revise constitutional structures whenever a citizen's preferences change in such a way that existing structures no longer satisfy a principle of equal degree of goal realisation – or whenever people enter or leave the society – it makes sense for contractarian agents to distribute equally the means of realising goals.[73] Each citizen has a direct access to her own dynamic preferences which others lack; placing control of the realisation of the goals associated with those preferences in the citizen's own hands allows her to respond more immediately and accurately to changes in her preferences. Equally distributing the means of realising goals, however, will amount to distributing rights, freedoms, and control over material resources in a manner compatible with liberal egalitarian principles.[74]

---

[72] Compare Rawls, 1999a: 132.

[73] See also §4.2, and the discussion of practical instability in §3.2.2.

[74] There is an obvious affinity here with Rawls's second moral power and his concept of primary goods (2005: 19, 74-75, 106; 2001: 19; compare 1999a: 442). I do not wish to claim that agents must be characterised as having a highest-order interest in *maintaining* the capacity to revise their ends – this builds a sectarian conception of the person into one's model. Agents are not irrational if they prefer to exercise their capacity in such a way as to surrender it. Nonetheless, given what one knows about oneself, and how one's ends have changed and developed over time, it will strike many people as sensible to allow for the fact that their present preferences – even those elements to which they are deeply committed – may turn out to be to some extent dynamic. Thus, while the contractarian theory may allow citizens to waive (or even alienate) more rights than many existing liberal states allow, there is no obvious reason to think the theory won't ascribe those rights to citizens in the first place. Note also that this approach does not presuppose an overly individualistic account

Third, among the available means of realising goals will be the exercise of state authority, and rational agents will seek both to secure for themselves as great a share of state authority as possible and to protect themselves from the exercise of state authority by others. The former will manifest as equal rights of political participation, and thus as some form of democracy.[75] The latter will manifest as constitutional limits on the operation of normal majoritarian procedures in the democracy. The balance between the two will depend on the marginal value each citizen attaches to being able to exercise authority over others as opposed to being immune from state authority in a particular domain.[76] Thus, for example, a citizen might care more about being able to practise her religion free from state regulation than about being similarly free to enter into contracts at any wage level she pleases. She might then wish the former to be subject to constitutional restrictions, while the latter is regulated by ordinary majoritarian means. Each citizen has the opportunity to impose such constitutional restrictions in the hypothetical choice situation, up to the

---

of citizens' conceptions of the good; those conceptions of the good which require collective action can be facilitated by providing citizens with the means by which they can voluntarily combine in associations for the pursuit of their ends (Rawls, 1999a: 249-251).

[75] Compare Christiano, 1996: 62-69. One might object that there is no requirement that the contractarian procedure issue in *equal* rights of political participation. Equality of degree of goal realisation is a principle which applies in the aggregate, rather than to each domain by turns. Agents will attach different values to participation rights, and so the contractarian procedure could allocate them unequal participation rights. There are three responses to this objection. First, so long as minimal participation rights are secured, the resulting system could still qualify as democratic (Saunders, 2010; compare Rawls, 1999a: 217). Second, as in the case of the basic liberal rights (n74, *supra*), citizens could be initially granted at least minimal participation rights, which they could later waive or alienate. The exercise of these powers of waiver or alienation could lead to an atypical form of liberal democratic society, but this possibility would not place the political institutions in question conceptually beyond the scope of liberal democracy, broadly conceived. Moreover, if this response is successful, there is no need to restrict ourselves to minimal participation rights – the combination of equal participation rights and robust powers of waiver and alienation is, from the point of view of the contractarian procedure, extensionally equivalent with unequal participation rights plus whatever compensating manner of goal realisation justifies the initial inequality. The only question is whether the bargaining which trades a share of state authority for the realisation of some other goal happens in the real world or in the hypothetical choice situation. Finally, it is worth noting that the dynamic preferences point is particularly pressing in the case of participation rights, since one never knows when one's fellow citizens will take it upon themselves to exercise state authority in a manner contrary to one's wishes. Retaining at least some capacity for democratic participation is a sensible precaution. Vanderschraaf (2019: 220-221, 223) also argues for democracy from a contractarian starting point, but in his account political equality depends dubiously on a Hobbes-style 'rough equality' assumption (see §1.6.2 and §1.6.3, *supra*). In this chapter and the next (*infra*), I aim to put the contractarian argument for democracy on a more secure foundation.

[76] Again bearing in mind my remarks about dynamic preferences, according to which citizens will take into account the fact that their preferences with respect to this balance may change over time.

limit where, by doing so, she attempts to claim a greater than equal share of degree of goal realisation. The precise constitutional form of a liberal democratic society is thus a contingent matter on the contractarian account, ranging from majoritarian democracies to democracies with extensive constitutional rights and robust judicial review. But this is not a defect in the theory, since there is no reason to think that the United Kingdom and the United States are not both liberal democratic societies.[77]

These considerations lead me to conclude that, in the central case of a contemporary society characterised by disagreement, the contractarian model will recommend political institutions which allow individuals personal freedom of a sufficiently wide scope, and recognise, through equality of degree of goal realisation (as always, for goals governed by political preferences), that each individual is equally entitled to that freedom (acknowledging that each might choose to exercise her freedom in a manner which ultimately deprives her of it).[78] It remains to consider the third component of our conventional understanding of liberal political institutions: a principle of neutrality or toleration which prevents the state unduly promoting sectarian conceptions of the good.

As I have already noted, it is perfectly possible for the contractarian model to prescribe political institutions which are wildly intolerant – the Trumparchy case, for instance. However, the model only recommends such outcomes when citizens' preferences are uniform, and so there is no need for toleration. Since I am assuming a diverse society, intolerant institutions cannot emerge from the model. Consider the case of religious toleration. A minimum requirement of being able to realise one's

[77] See Rawls's discussion of the four-stage sequence (1999a: 171-176).

[78] In a comparable discussion, Moehler goes further that I have here, and argues that contractarians should also endorse a capitalist, free market economy (2018: 184-188). While I have indicated that freedoms of exchange and contract are likely to be features of a contractarian society (n74 and n75, *supra*), Moehler seems to underestimate the extent to which there will be agents possessing much greater destructive or predatory powers than productive powers (see §7.5, and n29 in particular, *supra*). Such agents would be disadvantaged if their economic position in society reflected only their productive powers, as I take it would be the case under a capitalist economy, and would have no reason to agree to such a disadvantaged position from a Hobbesian baseline. For this reason, contractarians should leave the content of economic institutions relatively open.

goal of practising one's religion will be that doing so is not prohibited by law. If a citizen attaches sufficient value to being able to practise her religion, it will be impossible to achieve equality of degree of goal realisation without granting her freedom of religion.[79] This point generalises to non-religious conceptions of the good. Legal toleration of existing conceptions of the good is thus one of the necessary conditions of achieving equality of degree of goal realisation. If we add the prospect of dynamic preferences, toleration of both existing and prospective conceptions of the good follows. This approach also circumscribes the limits of toleration – what is intolerable in a contractarian society is a practice which threatens the realisation of a goal held by some citizen which is of the same order of importance as the preference for freedom (rather than restriction) of religious worship held by a religious citizen. Conceptions of the good which threaten the basic rights and freedoms of citizens will be intolerable, by this measure, but the pursuit of conceptions of the good which do not threaten basic rights and freedoms will be tolerated.

Under conditions of pluralism, then, the contractarian model will result in political institutions committed to some form of toleration. However, it is more difficult to argue that robust neutrality – understood as a day to day constraint on state actions – would follow.[80] The contractarian model is consistent with constitutional forms in which there are established churches, along the lines of the United Kingdom or Denmark, for example, as well as with constitutions in which there is a strict separation between church and state. The exact form of the constitution will be contingent on the different conceptions of justice in the population. Robust neutrality would require a significant proportion of the population to be morally committed to liberalism.

Thus, in the pluralistic conditions of contemporary societies, the political institutions prescribed by the contractarian model will be recognisable as liberal according to

[79] Cases of religious persecution – in which intolerant citizens have preferences against the practice of a particular religion which rank as highly as the preferences practitioners have in favour of the practice of the religion – will be cases of pathological preferences. See §5.5.2.
[80] §4.2 n52.

our conventional understanding of that term. I have also argued that the model meets five of the six elements of an authority-based conception of liberalism. In order to show that it fully meets the sixth, I must argue that, in addition to being equal in rationality, information, endowment, and prospects, agents in the contractarian model are properly represented as having equal bargaining power.

Chapter 8

To this point, I have simply assumed that agents in the contractarian model have equal bargaining power. Here, I argue for a qualified form of that claim. However, instead of approaching the question directly, I begin by examining a different kind of problem. This problem – the coalition problem – arises when we move beyond the 2-player scenarios I have been using to illustrate the model, and consider strategic interactions involving more than two agents.

Any useful theory of political institutions must be applicable to cases involving many agents. As such, the coalition problem deserves attention in its own right. Solving the problem requires introducing a number of new concepts. It is these concepts which will also provide the basis for arguing that agents in the contractarian model have equal bargaining power. Hence, I take an oblique approach, looking at the coalition problem first, before turning to bargaining power.

I will not initially be able to give a very precise definition of what bargaining power is. In bargaining theory, the concept of 'power' can play a variety of roles.[1] Earlier, I noted that bargaining power is represented in the Nash bargaining solution as a series of exponents (λ, μ...) which modify those terms of the solution linked to each party to the bargain (L, M...).[2] This, however, makes bargaining power a mathematical black box – we are supposed to fill in the exponents using whatever values we need to properly model our scenario. In the procedural elaboration of the Nash solution, bargaining power is identified with the players' discount rates or patience.[3] However, conceiving of bargaining power in this manner presupposes a structure like the alternating offers game, and not all bargaining occurs within such

[1] Compare Penrose, 1946: 53-57; Nash, 1950: 155; Nash, 1953: 137; Banzhaf, 1965; Binmore, 1998: 78-82; Osborne, 2004: 465-473.
[2] §6.3.
[3] §6.4.

a structure. The difficulty, then, is to state an understanding of bargaining power which does not prejudge the kind of scenario we are studying.[4]

For now, I will say that bargaining power is any feature of an agent's capacities or situation which (i) leads to her achieving a higher (expected) vNM payoff than she otherwise would and (ii) is not captured by some other part of the characterisation of her or her circumstances (her rationality, information, prospects, etc.). As a definition, this is vague and negative, but it has the advantage of flexibility. It will allow bargaining power to include things like patience or discount rates, but also to include asymmetries in agents' options (their capacity to make threats, for instance, or their capacity to join more than one possible coalition).

In §8.1, I introduce the coalition problem. In §8.2, I consider three unsuccessful responses to it. In §8.3, I offer an alternative response: the side payments approach. In §8.4, I defend the conception of rationality upon which the side payments approach relies. Finally, in §8.5, I use the side payments approach to argue for (duly restricted) equality of bargaining power in the contractarian model.

§8.1 The coalition problem

The *coalition problem* is a problem in bargaining theory.[5] To illustrate, consider a divide-the-dollar game, **[Game 1]**. In this game, there are three players: *A*, *B*, *C*. The

[4] Thoma (2015: 3336, 3347) presses this point further. If a mathematical representation of bargaining power is something which is rationally contestable, no bargaining solution relying on such a representation can be used by agents in conditions of profound pluralism.

[5] The coalition problem has received relatively little critical attention from political philosophers. In *Morals By Agreement*, Gauthier restricts his attention to 2-player cases in which the problem does not arise (1986: 140-141, 152-154, 190-191). Binmore, similarly, sets the problem aside (1994: 68, 75-78; 1998: 14-15), as does Moehler (2018: 108-109). The only explicit engagement with the problem of which I am aware is an exchange between Gauthier (1988) and Jean Hampton, but the problem is related to the equilibrium selection problem analysed by conventionalists such as Vanderschraaf (2019: 137-145, 155-159, 167-184; see also Gaus, 2018; and §1.6.1 and §1.6.3, *supra*). Conventionalists are generally interested in how a (game-theoretically) stable equilibrium solution to the coalition problem might evolve via uncoordinated actions by individual agents over repeated interactions, whereas the solutions I consider will focus on coordinated responses to the problem (see n11, *infra*).

object of the game is to end up with as much money as possible. The total prize available is a dollar, and the game ends whenever a majority of the players (so, either any two or all three) agrees on a division of the dollar.[6] In order to get a share of the dollar, a player must coordinate with at least one other player – she must form a coalition. There are four possible winning coalitions: *AB*, *AC*, *BC*, and *ABC*.

In the 2-player scenarios I examined earlier, a successful bargaining theory was defined by the fact that it could tell us what proportion of the coordination surplus the players (prudentially-)ought to demand from (and concede to) one another.[7] In **[Game 1]**, a bargaining theory must do more. Prescribing a division of the surplus requires telling us which of the possible coalitions (prudentially-)ought to divide the dollar. The difficulty is then that the game lacks what is called a *core*.[8] The core of a game is the set of stable winning coalitions. In games which lack a core, no winning coalition is stable. On standard accounts of rational choice, the absence of a stable winning coalition means that prudential reasons are insufficient to determine a choice of coalition, and thus the bargaining theory fails to prescribe any particular division of the available surplus.

To understand the problem, consider how **[Game 1]** might play out. Whenever any player proposes some particular division of the dollar, there is an alternative division which is preferred by at least two players. So, for example, if *A* proposes to split the dollar equally with *B* ($0.50, $0.50, $0.00), *C* can propose to *B* that they split the dollar instead, on terms more favourable to *B* than *A* offered ($0.00, $0.60, $0.40). *Any* 2-player winning coalition can be disrupted and displaced by an alternative winning coalition; and an excluded player, who will otherwise get nothing, will always have an incentive to make a disruptive offer. The same is true for the 'grand'

[6] A division is expressed (P_A, P_B, P_C...), where P_A is the proportion of the total prize, *P*, allocated to *A*, and so on for additional players. If no majority can agree on a division, each player will get nothing. In this, and in the other games discussed below, *P* is assumed to be perfectly divisible; the players are assumed to be rational and informed, and to strictly prefer greater shares of *P* to lesser shares; utilities are assumed to be linear over all proportions of *P*; there is no time limit on the game, and the players have no time-preferences with respect to *P* (that is, they don't care whether they get a share of *P* sooner rather than later; see Thoma, 2015: 3348).
[7] §6.1.
[8] Riker, 1982: 287; Hargreaves Heap *et al.*, 1992: 302-304; Binmore, 1998: 39-40, 68.

coalition, *ABC*. Suppose one of the players proposes an equal division of the prize, *P*: (⅓*P*, ⅓*P*, ⅓*P*). Both *A* and *B* can do better than this by excluding *C*.[9] Since each player's goal is to maximise the proportion of *P* which he or she receives, there is an incentive to reject the grand coalition in favour of a coalition with as few members as possible, and so the grand coalition is also unstable. Thus, for every proposed winning coalition, and every proposed division, there exists an alternative coalition and division which is superior from the point of view of at least two players. Any proposed bargain can be disrupted, and the cycle of offer and counteroffer has no apparent terminus. Importantly, we can generalise from this simple game to *n* dimensions and *m* parties.[10]

This conclusion has potentially disastrous implications for contractarian theories. The prescriptive force of a contractarian theory comes from the idea that you and I (prudentially-)ought to do what a rational agent would do. In games without a core, the structure of the bargaining situation is such that the theory is apparently unable to give a determinate account of how a rational agent would choose. If the coalition problem cannot be solved, contractarian theories will lack prescriptive content whenever the precontractual bargaining situation lacks a core.

§8.2 Unsuccessful responses to the coalition problem

§8.2.1 The pragmatic response

Faced with the coalition problem, one might be inclined to make a *pragmatic* response. We know that, for instance, parliamentary coalitions can form and hold in the real world, despite the fact that a formal description of the bargaining situation would often characterise the games as lacking a core. Our pragmatist will say that, at some point, even in games without a core, agents will come to an agreement in

[9] As can the other possible pairs, *AC* and *BC*.
[10] This slightly overstates the case. There are conditions under which some such systems would be in equilibrium (Riker, 1982: 185). For the theoretical underpinnings of this approach to coalitions see McKelvey (1979).

order to divide the prize. The contractarian is supposed to be modelling real people, and real people do not sit around bargaining indefinitely. Particular circumstances – the structure of the political party system or the institutional context within which bargaining occurs – can play an important role in determining the outcome of coalition bargaining. By whatever means they escape it, real people are not troubled by the coalition problem.

This response needs to be evaluated with care. The first caveat is that the conclusion is too strong. It's true that the coalition problem is less important in practice than it is in idealised bargaining theories, but it is an exaggeration to say that real political actors are never concerned with it at all. Political systems do go through periods of shifting coalitions and instability. That said, the basic point stands. If real people can form and sustain coalitions in strategic circumstances which lack a core, then the coalition problem looks like a quirky artefact of an overly abstract theory.

Unfortunately, this pragmatic response is inadequate. Contractarianism is a means of testing our existing institutions by imagining ourselves to be free to choose whatever institutions we like and then asking if we have good reasons to choose existing ones (and if we don't, what set of alternatives, if any, we have decisively good reasons to choose). As a matter of empirical fact, it may well be the case that historical, institutional, and wholly arbitrary constraints actually determine which coalitions form in the real world. That is a question for political science, history, and social psychology. A contractarian theory, however, is supposed to be providing us with *reasons* to endorse one division of the benefits and burdens of coordination over another.

The pragmatic response says that the coalition problem is not a problem because, in the real world, exogenous factors will resolve the cycle of bargaining in one way rather than another. This response misses the point: the theory is supposed to be determining the outcome of bargaining *endogenously*.[11] If it turns out that the

[11] Endogeneity is the condition with respect to which conventionalist accounts are suspect, to the extent, say, that the equilibrium points upon which they converge are problematically path

outcome of the bargaining upon which contractarian theories rely is not determined by the weight of prudential reasons, but is essentially arbitrary, then we lack a reason to endorse the institutions that emerge from any such bargain.[12]

§8.2.2 The natural core response

As I have set it out, the coalition problem arises when a bargaining situation lacks a core. The most straightforward way to resolve the problem might then be to show that the contractarian bargaining situation necessarily has a core. Call this the *natural core* response.

I doubt that a plausible specification of the contractarian choice situation can be given which guarantees that the game will have a core. But, for the sake of argument, suppose the situation does indeed have a core. Even so, a further problem arises: When we look at certain games, we find that solutions within the core do not exhaust the apparently rational options available to the players. If it is rational to select a coalition outside the core, then the existence of a core does not resolve the coalition problem. While there will be stable winning coalitions available, it will not be rational to select them.

To see how this can be the case, I need to look in detail at two examples discussed by Gauthier. We have met one already: Sam McGee is a cash-strapped prospector who wishes to register a claim to a vein of gold he has discovered; Grasp, a banker, is the only person who can lend McGee the money he needs.[13] In the second example, Ms Macquarrie, a pharmaceutical chemist, "requires a laboratory assistant

dependent (compare Thrasher, 2014: 690-691; Moehler, 2018: 109-110, 121-125, 147-148, 154-155; Thrasher, 2019: 446-447). See n5, *supra*.

[12] Thoma, 2015: 3336, 3343-3345. The fact that an actual bargain *has been struck* – a contract or promise made – may be a good reason to *stick* to the bargain. This is a separate concern, however. The contractarian theory is supposed to be telling us what bargain we (prudentially-)ought to have made, not whether or not we (in-some-sense-)ought to stick to whatever bargain we have made. This latter is an interesting question, but it is not my concern here. See the discussion of arbitrariness and determinacy in §1.6.1, *supra*. Compare Vanderschraaf, 2019: 45, 48-49, 52, 217-218.

[13] §6.4.

to aid her in carrying out her experiments, and hires Mr O'Rourke."[14] As a result of the experiments, a valuable new drug is developed, promising to make those who produce it very rich.

According to Gauthier's bargaining theory, Grasp may rationally demand half the value of McGee's claim. By contrast, Macquarrie pays O'Rourke *x*, the going rate for lab assistants (*x* being much less than half the proceeds from the new drug). Although Gauthier concedes that Macquarrie would be "ungenerous not to give O'Rourke a handsome bonus" once the project pays off, "she does not owe him anything."[15]

The reason Gauthier prescribes an equal division in the McGee-Grasp case, but not in the Macquarrie-O'Rourke case, is that Grasp's provision of the loan is *essential* to McGee's registration of his claim, whereas, by hypothesis, O'Rourke is one of a number of equally well-qualified research assistants Macquarrie might have chosen to hire. Grasp can credibly threaten to deny McGee any share of the value of the claim by withholding his coordination. O'Rourke is not in a position to similarly threaten Macquarrie. If O'Rourke rejects Macquarrie's offer of *x*, she will simply hire a different assistant. Grasp and O'Rourke thus have different degrees of bargaining power. Grasp's bargaining power secures for him half the value of the claim; O'Rourke's lack of bargaining power relative to Macquarrie means he has to settle for a meagre wage.

Now, consider a variation in the Macquarrie-O'Rourke example: **[Game 2]**.[16] Suppose that, as before, Macquarrie needs to hire an assistant, but there are only two assistants who are qualified to work on her project. Mr O'Rourke is one; the

[14] Gauthier, 1986: 153.
[15] *Ibid*.
[16] My extension of the Macquarrie-O'Rourke example closely follows Gauthier's discussion of Abel, Gable, and Mabel (1988: 396-397). The discussion in the text draws on my MPhil thesis (2006). This version of the example is obviously somewhat contrived, since it initially takes *x* to be a market rate that is presumably the result of many research assistants competing for many jobs but also assumes that there are only two research assistants, O'Rourke and O'Reilly, in the market for this particular job. Despite this contrivance, playing out the example will be helpful in understanding Gauthier's rejection of the natural core response.

second is Ms O'Reilly. The structure of the game is such that a coalition of Macquarrie and either one of the assistants can produce P (the proceeds from the new drug). The addition of a third player does not increase the surplus (Macquarrie only needs one assistant to complete the project; having two doesn't make the drug any more valuable). Finally, the assistants by themselves cannot produce anything (Macquarrie is the only chemist with the knowledge required to produce the drug).

Because neither of the assistants is essential to the production of the surplus, a straightforward application of Gauthier's bargaining theory awards the whole of the surplus to Macquarrie (less the x which is assumed to be the marginal rate at which both O'Rourke and O'Reilly are indifferent between work and leisure). Suppose that Macquarrie offers this division, $(P - x, x, 0)$, to O'Rourke. Recall that, on Gauthier's account of rational concession, a party who stands to lose more by a failure to come to agreement is rationally required to concede to another.[17] At this point, the cost to O'Rourke of failing to reach agreement is x (the offer he is leaving on the table by turning Macquarrie down). Crucially, the cost to Macquarrie is 0, since if O'Rourke refuses the offer, Macquarrie will make the same offer to O'Reilly and (assuming O'Reilly accepts) will receive $(P - x)$ without O'Rourke's coordination. Since $x > 0$, any further concession must be made by O'Rourke not Macquarrie. The only reason that O'Rourke doesn't make a further concession is that x is the lowest payoff which would make coordination worth his time. Thus, the straightforward application of Gauthier's bargaining theory prescribes a $(P - x, x, 0)$ division between Macquarrie, O'Rourke, and O'Reilly.

This division, and the division in which O'Reilly is substituted for O'Rourke, are the only solutions in the core of **[Game 2]**. Any proposal to give an assistant less than x would be insufficient to induce either O'Rourke or O'Reilly to join the coalition, since each can get $(x - \varepsilon)$ on his or her own (this is the value of leisure for the duration of the project, and we assume that O'Rourke and O'Reilly attach the same marginal value to leisure). Any proposal to give the assistant more than x would be unstable,

[17] §6.4.

because two players would always prefer an alternative division. That is, if it is proposed to give O'Rourke ($x + \varepsilon$) instead of x, both Macquarrie and O'Reilly will prefer to form a coalition in which Macquarrie only has to surrender x to O'Reilly.

So, solutions in which the assistant receives x are the only solutions in the core, and a straightforward application of Gauthier's bargaining theory selects those solutions. But, Gauthier says, this does not appear to be rational or, for what it's worth, fair.[18] If O'Rourke and O'Reilly were able to cooperate together and negotiate jointly with Macquarrie, they would be able to secure for themselves a much better deal.

Suppose that, as before, Macquarrie makes a job offer to O'Rourke at x, the market rate; if O'Rourke rejects the offer, Macquarrie will offer the same rate to O'Reilly. O'Rourke and O'Reilly discuss their situation and come to an agreement. For their mutual benefit, they will pool their resources and adopt a joint negotiating strategy with respect to their dealings with Macquarrie.

Since, combined, the O'Rourke-O'Reilly partnership is essential to Macquarrie's project, it is in a position equivalent to the one Grasp occupies in Gauthier's tale of Sam McGee. The partnership may demand half of P in negotiations with Macquarrie and rationally expect Macquarrie to concede, because now the cost to Macquarrie of failing to come to an agreement is $\frac{1}{2}P$. Once the project has been completed, O'Rourke and O'Reilly will divide their share of P equally between them.[19] Without the partnership, O'Rourke would have been forced to accept x. So long as $\frac{1}{4}P$ (half of $\frac{1}{2}P$) is greater than x, the partnership is rationally defensible (and O'Reilly's position is symmetric). A sophisticated application of Gauthier's bargaining theory thus divides the surplus "through two agreements, one involving [O'Rourke and O'Reilly], the other involving, in effect, their coalition and [Macquarrie]… The final outcome would not be in the core of the three-person interaction, but each part

[18] 1988: 397. Compare Riker and Ordeshook, 1973: 140.

[19] This is a straightforward application of either the Nash solution or the Kalai-Smorodinsky solution to the O'Rourke-O'Reilly sub-game. To be more precise, only one assistant will be needed to work on the project, so some kind of side deal will be required between O'Rourke and O'Reilly to either share the work or compensate for the cost.

would fall in the core of one of the two-person interactions… into which the original interaction has been decomposed."[20]

Thus, even though a core existed for the 3-player game, it is unacceptable to restrict rational bargaining to solutions within that core. Surely O'Rourke and O'Reilly do not act irrationally in coordinating to maximise the share of the surplus which they receive, and if so, rationally mandated solutions cannot be limited to the core. Even if it were possible to show that all plausible characterisations of the contractarian choice situation would describe it as a game with a core – and, again, I doubt this – the Macquarrie-O'Rourke-O'Reilly game demonstrates that solutions in the core are not necessarily adequate bargaining solutions. It can be rational to form coalitions to move away from the core, and once we move to a solution outside the core, the coalition problem reemerges. The existence of a core does not guarantee an escape from the coalition problem, and so the natural core response to the problem fails.

§8.2.3 The cooperative infrastructures response

The Macquarrie-O'Rourke-O'Reilly game provides an outline of how Gauthier might go about extending his theory to incorporate a solution to the coalition problem.[21] On Gauthier's view, all *n*-player games can be decomposed into a sequence of 2-player games. At each point in the sequence, we form a larger coalition from the combination of two agents, where an agent can be either an individual or a coalition formed at an earlier point in the sequence. Since the coalition problem only arises in games with more than two players, this solution avoids the problem by reducing all *n*-player games to a collection of 2-player games and then dealing with each of the 2-player games in turn.

[20] Gauthier, 1988: 397.

[21] I say *might* go about because there is a significant amount of inference required here. Gauthier's discussion (*ibid*.: 395-398) is not intended to be comprehensive, and it would be unreasonable to understand it as his considered view, even assuming I am interpreting it correctly. Nevertheless, the discussion is suggestive, so I explore its implications.

In games with more than two players, there will always be more than one possible order in which to go about combining agents into coalitions. Each of these orderings is a different *cooperative infrastructure*. The challenge for Gauthier is to identify a particular infrastructure as the one agents (prudentially-)ought to employ, and to do so in a way that avoids both infinite cycling over possible coalitions and exogenous determination of a winning coalition. I will argue that, in general, there is no way this can be done.

The cooperative infrastructures response to the coalition problem fails in the face of either of two objections. Gauthier is aware of the first, but seems to think it doesn't invalidate the approach.[22] The first objection asks if Macquarrie can make any proposals that can disrupt the O'Rourke-O'Reilly coalition. It seems that she can. If O'Rourke and O'Reilly are each getting $¼P$ from the partnership arrangement, then any offer greater than $¼P$ should induce one of them to leave the partnership. Thus, suppose Macquarrie makes an offer of y to, say, O'Reilly ($¼P < y < ½P$). Faced with the prospect of getting nothing at all, and seeing that O'Reilly is ready to cut a deal, O'Rourke counters with an offer to Macquarrie of $¼P$ (which is, of course, less than y). Now O'Reilly stands to gain nothing at all; as a consequence, she makes an offer to Macquarrie at $¼P$ less some ε. O'Rourke and O'Reilly will continue to undercut one another's bids until they reach x, the marginal rate at which working for Macquarrie is preferred to not working at all. O'Rourke and O'Reilly are right back where they started.

It may be thought that this is too quick – all I've done is rehearse the problem to which Gauthier is providing a solution; I haven't undermined that solution. Rational agents can see the regress coming and react accordingly. If there were some way in which players could resolutely lock themselves into coalitions, then O'Rourke and O'Reilly could commit to the partnership, rendering themselves rationally incapable of defecting in the face of Macquarrie's offers. And it looks like Gauthier has a plausible candidate available to fill this role – constrained maximisation.[23]

[22] *Ibid*.: 397 n8, 398.
[23] 1986: 167-170; compare Moehler, 2018: 103-107. See also the Introduction and §7.5, *supra*.

Constrained maximisers can resist incentives to defect from agreements when doing so is required to realise greater strategic goals (in this case, achieving a greater share of P).

Unfortunately, this does not solve the problem, because we now have to ask how Macquarrie can best respond to a strategic environment in which the players can make binding commitments. Macquarrie can foresee that O'Rourke and O'Reilly will form a partnership, and can preempt it by making an initial offer of y to, say, O'Reilly, in order to lock O'Reilly into a partnership with her instead of with O'Rourke (again, $\frac{1}{4}P < y < \frac{1}{2}P$). In order to tempt O'Reilly into a binding coalition with him, O'Rourke would have to offer to accept $\frac{1}{2}P - y$, which is less than $\frac{1}{4}P$. Instead, he could go to Macquarrie and point out that if he enters a coalition with O'Reilly, Macquarrie will be forced to hand over $\frac{1}{2}P$; whereas, if Macquarrie offers him $\frac{1}{4}P$, which is less than she is proposing to give to O'Reilly, he will accept it rather than enter the coalition with O'Reilly. O'Reilly can respond either by undercutting O'Rourke, in which case the regress has begun once again; or by proposing that, in order to avoid the regress, O'Rourke should enter a binding coalition with her, in which case Macquarrie can make the same preemptive offer of y to O'Rourke that she initially made to O'Reilly (and the sequence of offers and counteroffers plays out as before, with O'Rourke substituted for O'Reilly). Introducing the possibility of resolute coalitions has simply shifted the coalition problem back a step.

This, then, is the second objection to Gauthier's multi-stage solution. Suppose we abandon the particular details of the Macquarrie-O'Rourke-O'Reilly example and return to a version of our divide-the-dollar game: In **[Game 3]**, solo players receive nothing and 2-player coalitions may divide P. However, this time, the grand coalition may divide $2P$. Suppose that A and B form a coalition in stage one of the game – they agree to share any proceeds equally; each will get $\frac{1}{2}P$, plus half of the coalition's share of any agreement with C. The coalition then makes an offer to C – the proposal is to divide the payoff of coordination equally between the coalition and C. The snag is that the AB coalition can secure P on its own, so the surplus to

be divided between the coalition, *AB*, and *C* is just *P* (that proportion of the total, 2*P*, on which the coordination of *C* depends). *C*, therefore, gets a payoff of ½*P*. *A* and *B* each get a payoff of ¾*P*. Being able to form a coalition in the first stage of the game has increased the payoff received by the members of the initial coalition at the expense of the player who was initially excluded.

Note, of course, that nothing in the description of the game requires that stage one involve an agreement between *A* and *B*, rather than between, say, *B* and *C*. By substitution, we can show that *B* and *C* would benefit, had we begun with a *BC* coalition. The outcome of the game varies according to the choice of initial coalition, so we need some mechanism for selecting an initial coalition. If our solution to the coalition problem involves decomposing an *n*-player game into a sequence of 2-player games, and if the various ways in which we can combine the *n* players will lead to different allocations of the payoff, then we need some ordering principle for selecting one cooperative infrastructure from among the various options.

Are there any ways to determine the order in which coalitions should form? In answering this question, the immediate danger that must be avoided is exogenous determination.[24] Since the order of coalition formation determines the outcome of the game, if the order is determined by factors other than the rational pursuit by the players of their ends, then the division of *P* ultimately prescribed is determined exogenously. If applying the cooperative infrastructures solution entails that the outcome of the game is determined exogenously, then any contractarian theory relying on this solution would lack prescriptive force.

Although Gauthier does not discuss principles for selecting a cooperative infrastructure, the most natural move (and, as far as I can see, the only one which will avoid the arbitrariness objection) is to argue that a rational player would choose the order of coalition formation that will maximise his or her payoff. In **[Game 4]**,

[24] §8.2.1.

for instance, there are three players, and hence three options available for the choice of initial coalition: *AB*, *AC*, and *BC*. Payoffs for the possible 2-player and 3-player coalitions vary: Any solo player receives a payoff of 0. *AB* can secure *P*; *AC* can secure 2*P*; and *BC* can secure 3*P*. The grand coalition secures 4*P*.

The ultimate payoffs to each player, for each choice of initial coalition, are as follows:

	A	*B*	*C*
Initial *AB*	$5(\frac{P}{4})$	$5(\frac{P}{4})$	$6(\frac{P}{4})$
Initial *AC*	$6(\frac{P}{4})$	$4(\frac{P}{4})$	$6(\frac{P}{4})$
Initial *BC*	$2(\frac{P}{4})$	$7(\frac{P}{4})$	$7(\frac{P}{4})$

Table 1[25]

On this basis, *A* most prefers the initial coalition *AC*, followed by the initial coalition *AB*, and ranks the initial coalition *BC* as her least preferred option. By contrast, both *B* and *C* rank the initial coalition *BC* as their first preference. Given that *B* and *C* both prefer an order of coalition formation in which *BC* is formed in the first stage of the game, and that the formation of this coalition does not require the agreement of *A*, we seem to have a non-arbitrary basis for selecting a cooperative infrastructure, which is what Gauthier's solution needs.

[25] In Table 1, for convenience, I have expressed all payoffs in units of ¼*P*. The entry in the upper-left cell is the ultimate payoff to *A*, given an initial coalition *AB*, and so on through the rest of the table. There will be 16 units to allocate in total (there are 16 ¼*P*s in 4*P*). If *AB* is the initial coalition, *AB* can secure *P* by itself, so the surplus for division between *AB* and *C* is 12 units. The bargaining solution allocates 10 units to *AB* (*P*, plus half of 3*P*) and 6 units to *C* (half of 3*P*). *AB* then splits its 10 units equally, giving *A* and *B* 5 units each, for an ultimate division of (5, 5, 6). The other rows are filled in similarly, starting with a different initial coalition.

Two further problems now arise. First, the success of Gauthier's solution in **[Game 4]** relies on there being one order of coalition formation which dominates all alternatives. Selecting an order for coalition formation, however, is a social choice. As with any social choice, there is the possibility of a Condorcet cycle arising from the preferences of the agents. Whenever such a cycle is present, this approach cannot resolve the second-order bargaining game which is needed to select a cooperative infrastructure.[26]

It may be, though, that such cycles are not very common. So long as the cooperative infrastructures approach can avoid collapsing into cycling in typical cases it remains a viable solution to the coalition problem. For the sake of argument, let's assume that **[Game 4]** represents such a typical case. Is the cooperative infrastructures approach secure?

Unfortunately not. *A* is the player most disadvantaged by the formation of the coalition *BC* in the initial phase of the game; if she cannot disrupt the coalition, she will achieve her least-preferred outcome (2 units). Just as in the Macquarrie-O'Rourke-O'Reilly game, we need to ask if *A* can make a disruptive offer. In order to disrupt the coalition, *A* has to make an offer to either *B* or *C* that will ultimately give one of them more than he will secure by initially forming the *BC* coalition (7 units).

Let's look at the possibility of making a disruptive offer to *B*. The coalition *AB* has 10 units to distribute between the members. When *A* is excluded from the initial coalition, the sum of what *A* receives (2 units) and what *B* receives (7 units) is less than the payoff that the coalition *AB* has to distribute between the coalition partners (10 units). As such, there is some division of that payoff which is superior to Gauthier's solution from the point of view of both *A* and *B*, and therefore *A* can make an offer that disrupts the coalition *BC*.[27] Similarly, *A* can disrupt the coalition

[26] This is the case in **[Game 3]**, for instance. Compare Thrasher, 2019: 440-441.
[27] If *A* proposes to give *B* 7.5 units, and keep 2.5 units for himself, such a division would be superior for both players.

by making an offer to *C* which falls within a certain range.

Even though there is a dominant order of coalition formation, the game still isn't stable. *A* has the power to make a disruptive offer, and every incentive to do so. And, once *B* has switched to supporting *AB* as the initial coalition, *C* will be able to make a disruptive offer to either *A* or *B*. Again, we're forced back into a bargaining cycle, and the cooperative infrastructures approach has failed to provide us with a solution to the coalition problem.

The success of any solution to the coalition problem relies on preventing the emergence of a cycle of offers without appealing to exogenous factors. Gauthier's cooperative infrastructures approach does prevent the cycle, assuming an order of coalition formation is given, by limiting the number of players at each stage of the game to two. The difficulty arises in identifying the order – in selecting one cooperative infrastructure from the many available. Given the possibility that the outcome of a game varies according to the infrastructure selected, the players are not indifferent to the order in which coalitions are formed. Given that the players must coordinate in order to impose a cooperative infrastructure on the game, the selection of an infrastructure becomes a second-order bargaining problem, one which the cooperative infrastructures approach is itself ill-equipped to handle.[28]

In order to avoid an interminable regress, it must be the case, first, that there is an order of coalition formation which is preferred by all the players necessary to implement it, and second, that players disadvantaged by the dominant infrastructure cannot make offers which disrupt the selection of that infrastructure. Absent good reasons to think these conditions hold for typical contractarian choice situations, the cooperative infrastructures approach does not offer a viable general solution to the coalition problem.

[28] Compare Thoma, 2015: 3349.

§8.3 The side payments response

The responses considered thus far do not succeed in solving the coalition problem. Still, it remains the case that some solution is essential for a contractarian theory. In this section, I propose the *side payments* response.[29] I argue that this response successfully avoids both infinite cycling and exogenous determination.

§8.3.1 Bargaining power as a function of strategic power

Let me return to the McGee-Grasp example.[30] Recall that McGee wished to register a claim to a vein of gold that he had discovered and that, because Grasp was the only person available to lend him the $100 needed to do so, Grasp could rationally demand half the value of the claim. This is true despite the fact that Grasp contributed just $100 to the joint enterprise of registering the claim, while McGee may have spent a lifetime searching for the gold. To emphasise the point, we can say that the value of the vein of gold is $1 million, and the money necessary to register the claim is just $1. So long as it is the case that McGee needs Grasp's coordination to register his claim to the gold, both the Nash and Kalai-Smorodinsky bargaining solutions hold that Grasp may rationally demand (and rationally expect McGee to concede) half the value of the claim. The extent to which each party contributes in monetary terms, or in terms of effort, to the realisation of the joint goal is irrelevant.[31] What matters is the strategic position in which the participants find themselves – the extent to which each participant is reliant on the coordination of the other to access the prize, and thus the extent to which each participant is vulnerable to having her access to the prize blocked by the other.

I said earlier that bargaining power is any feature of an agent's capacities or situation which (i) leads to her achieving a higher (expected) vNM payoff than she otherwise

[29] For a suggestion which partially anticipates this response, albeit in a different context, see Riker and Ordeshook (1973: 140).
[30] §6.4, §8.2.2.
[31] Gauthier, 1988: 390-394.

would and (ii) is not captured by some other part of the characterisation of her or her circumstances.[32] In the McGee-Grasp case, the participants' bargaining power is determined by their strategic positions. In this particular case, the strategic positions of the participants are symmetric, and so we can say that McGee and Grasp have equal strategic power – each has the option to veto the bargain, leaving the other with nothing. The participants' bargaining power consists in their capacity to exercise their strategic power, or to threaten to exercise it. As a result of having equal strategic power, the participants have equal bargaining power, and so they divide the payoff of the game equally.

Now, consider **[Game 1]** again. As before, any two players can secure *P* (a dollar), and the grand coalition secures no additional payoff. This is the classic game without a core. Suppose that *A* proposes to form a coalition with *B* to secure *P*. Nonetheless, suppose *A* is what we can call *globally rational*, as opposed to merely *locally rational*.[33] She realises that *C*, the excluded player, occupies a certain strategic position in the game. *C* can, through making an offer to *B* that undercuts *A*'s offer, begin the process of bid and counter-bid that leads to the infinite cycle of 2-player coalitions. Just as McGee has an interest in securing Grasp's coordination, *A* has an interest in the actions of *both B and C* – she wants *B* to accept her offer, and she wants *C* to refrain from disrupting it.

In order to induce *B* to enter a coalition, *A* must offer him some share of the dollar – call this *y*. Because *A* is globally rational, and can anticipate the coming cycle, there now arises for *A* the question of how to induce *C* to refrain from disrupting the *AB* coalition. All *C* cares about, and all *A* has to offer, is (some share of) the dollar. So, in order to induce *C* to refrain from disrupting the coalition, *A* offers to make a side payment to *C* – call this *z*. Having proposed to give *y* to *B*, and *z* to *C*, *A* then proposes to keep the remainder of the dollar for herself – call this *x*. Thus, $x + y + z = P$ (in this case, \$1).

[32] See the introductory remarks for this chapter.
[33] On this distinction, and distinctions which get at similar ideas, see Parfit (1984: 497-499), Hansson (1993: 484), Sobel (1997: 42-57, 67-68), Rabinowicz (2000), and Aldred (2007: 402).

What are the values of x, y, and z? Following the intuitive logic of the McGee-Grasp case, it seems it must be the case that players whose situation is strategically identical must be treated the same. As all players in the divide-the-dollar game have the same strategic powers – the same capacity to form coalitions, and to make disruptive offers – this suggests that $x = y = z = ⅓$ of the dollar.[34]

This, then, is the side payments response to the coalition problem. It suggests that bargaining agents can rationally demand, and rationally expect that other agents will concede, a share of the surplus arising from coordination that is proportional to their bargaining power, where bargaining power is understood as a function of the strategic options available to them. In the context of a divide-the-dollar game, and of games of a similar form, the key capacities contributing to bargaining power are the capacity to form coalitions and the capacity to make offers which disrupt a proposed winning coalition. Any player who possesses this latter capacity must be offered a side payment to induce her to refrain from exercising it. In the particular case of **[Game 1]**, all players possess the capacity, and possess it in the same degree. Thus, the side payments response recommends an equal division of the surplus between the three players.

If the side payments response is successful, the coalition problem will be solved. Infinite cycling will be prevented, because any player with the capacity to make a disruptive offer will have a decisive reason to refrain from doing so. Exogenous determination will have been avoided, because the division of the surplus is determined only by factors internal to the game: the strategic positions of the agents, and their rational preferences.

[34] More formally, given that the players have been specified identically, if the values of x, y, and z were not equal, the procedure would violate a standard *anonymity* condition (May, 1952: 681).

§8.3.2 Justifying side payments

Given that the players in **[Game 1]** are functionally identical, an equal division of the surplus is intuitively plausible. What's less clear is the reasoning for the division. Our intuitions could be led to an equal division by a substantive ethical commitment to fairness, rather than by an instrumentally rational, amoral appreciation of the strategic logic, as I would like to claim.[35] If the side payments response is going to be consistent with the contractarian model, it cannot rely on an appeal to fairness.

To counter this worry, suppose there are four players (*A*, *B*, *C*, and *D*) in a weighted voting game, **[Game 5]**. Suppose the prize is *P*, and that it can be divided by any coalition attracting a majority of the votes in the game, where *A*, *B*, and *C* each have 3 votes, and *D* has 1 vote. In this game, the side payments response recommends a division (⅓*P*, ⅓*P*, ⅓*P*, 0). Any coalition involving at least two of players *A*, *B*, and *C* can realise *P*. Player *D* can be part of winning coalitions but cannot be a *necessary* part. As such, she cannot credibly threaten to leave one coalition in favour of another, and other players have no incentive to attempt to induce her to do so.[36] *D* is a dummy player – she has no strategic power in the game – and so the outcome of this game is a division of *P* that gives each player who does have strategic power a payoff of ⅓*P*. Contrast this with the suggestion that it would be 'fair' to divide the prize evenly among all four players, or 'fair' to divide it in proportion to the votes that each has. We can use cases like this to satisfy ourselves that the intuitive heavy-lifting isn't being done by fairness in the side payments approach.[37]

Similarly, consider how the side payments response would work in the Macquarrie-

[35] Compare Gauthier, 1988: 398-399.

[36] Any winning coalition which contains *D* necessarily must also contain at least two of *A*, *B*, and *C*. Any two of *A*, *B*, and *C* can realise *P*, so subtracting *D* from the winning coalition would not reduce the payoff for the coalition; it would, however, increase the payoff available for division between the other two or three members of the coalition. Thus, the 4-player game reduces to the 3-player divide-the-dollar game. To put this another way, *D* is not in the circumstances of justice with the other players (§1.5 n37).

[37] See also the discussion of the McGee-Grasp example in §8.3.1. In that case, the proportion of the prize allocated to Grasp by the side payments response is entirely independent of the size of his financial contribution, or his efforts. This again tells against the idea that any substantive conception of fairness is at work.

O'Rourke-O'Reilly game. Suppose that Macquarrie offers O'Rourke some $x << \frac{1}{2}P$ in return for his coordination. If he refuses, she intends to make the same offer to O'Reilly (recall that x is the marginal rate at which O'Rourke and O'Reilly will coordinate). O'Rourke thinks for a moment and makes the following counter-proposal. He will agree to coordinate with Macquarrie in return for a payoff of $\frac{1}{2}P$; recognising the fact that O'Reilly can undercut his bid, he proposes to make a side payment to her of some y, in order to induce her to refrain from disrupting his offer (or accepting a disruptive offer from Macquarrie); he proposes to keep z for himself ($y + z = \frac{1}{2}P$). Again, given that we assume O'Rourke and O'Reilly to occupy an identical strategic position, it seems that $y = z = \frac{1}{4}P$.

Two questions now arise. First, in the context of the Macquarrie-O'Rourke-O'Reilly game, why isn't the side payments response vulnerable to the regress of disruptive offers which undid the cooperative infrastructures response? Second, if the side payments response isn't undone in the same way, why doesn't it recommend an equal division of the prize in the Macquarrie-O'Rourke-O'Reilly game, as it did in the divide-the-dollar game?

Take the first of these questions, and assume that the state of the game is one in which O'Rourke has proposed a division ($\frac{1}{2}P$, $\frac{1}{4}P$, $\frac{1}{4}P$). Which of the players has a reason to make a disruptive offer? It is true that both Macquarrie and O'Reilly could profit from an alternative division. O'Reilly, however, is globally rational – she can anticipate that if either she or Macquarrie proposes a division which allocates more to either or both of them, this will mean allocating less to O'Rourke, and O'Rourke will then have a reason to make a disruptive offer which undercuts O'Reilly. The only way to block this disruptive offer is via a side payment, but any side payment would have to restore O'Rourke to $\frac{1}{4}P$, and this would eliminate the gains of the original disruptive offer. O'Reilly, therefore, cannot move away from O'Rourke's proposed division and also block a disruptive offer; and if she fails to block such an offer, the regress will result in her gaining $x << \frac{1}{4}P$. Thus, O'Reilly has no reason to

make a disruptive offer.[38] Macquarrie, on the other hand, does have a reason to make a disruptive offer. However, she lacks a partner who will accept one – both O'Rourke and O'Reilly can anticipate that the resulting regress will leave them worse off than under O'Rourke's proposal. Thus, given global rationality, none of the players can successfully disrupt the proposed division.

Turning to the second question, why does the side payments response not return an equal, three-way split in the Macquarrie-O'Rourke-O'Reilly game? At first, one might think it should. Suppose, for instance, that O'Reilly proposes (⅓*P*, ⅓*P*, ⅓*P*). According to Gauthier's account of bargaining, a player who stands to lose more by a failure to make a marginal concession is rationally required to concede.[39] In any (Pareto-optimal) division other than an equal one, some player appears to have more to lose from a deadlock than another.[40] For example, if the proposal is (½*P*, ¼*P*, ¼*P*), it appears Macquarrie stands to lose ½*P*, while the other players stand to lose ¼*P* each, and so Macquarrie must concede further. According to this reasoning, the only division under which no player has a reason to concede further is the equal division.

This line of argument goes astray because it fails to appreciate that the players in the game are not strategically equal. Macquarrie occupies a stronger position relative to the other players – she is a necessary member of *all* winning coalitions. Crucially, the relevant non-agreement point for Macquarrie is not the case where the project does not go ahead and everyone gets nothing. Rather, Macquarrie can at any time offer either assistant ½*P*; this offer will be accepted, and cannot be disrupted by the excluded player. When the proposal is (⅓*P*, ⅓*P*, ⅓*P*), Macquarrie stands to lose *less* from a failure to agree than the other two players, because she stands to lose *nothing*.

[38] For the purpose of illustration, I have assumed the offer on the table is from O'Rourke. However, if it were from another agent, O'Rourke's position would be symmetric with O'Reilly's. If she lacks a reason to make a disruptive offer when the proposed division is (½*P*, ¼*P*, ¼*P*), so does he.
[39] This is an application of Kalai-Smorodinsky. See §6.3 and §6.4. The same result follows from a naïve application of the Nash solution, in which the exponents have not been adjusted to take the potential coalition dynamics into account (see the introductory remarks for this chapter).
[40] I assume transferrable utility here (§6.4 n30).

To illustrate, suppose Macquarrie offers ½*P* to O'Rourke. In order to disrupt this offer, O'Reilly would have to offer O'Rourke more than ½*P*. Because Macquarrie is a necessary member of any winning coalition, the only way to allocate more than ½*P* to O'Rourke is for an O'Rourke-O'Reilly coalition to hold out for more than ½*P* in its negotiation with Macquarrie, and then have the gains diverted to O'Rourke. But, in such an event, O'Rourke would stand to lose more than Macquarrie from a deadlock (it cannot be the case that both Macquarrie and O'Rourke stand to lose more than ½*P*). In those circumstances, O'Rourke would be rationally required to concede in order to equalise his relative concession with Macquarrie's. And then neither O'Rourke nor Macquarrie has any reason to equalise their concessions at any point other than ½*P*, excluding O'Reilly.

By substitution, we can show that the same logic holds for an initial offer of ½*P* to O'Reilly, and so an offer by Macquarrie of ½*P* to one of the assistants cannot be disrupted by the other. Thus, at the relevant non-agreement point, Macquarrie receives ½*P*. Macquarrie, therefore, cannot fail to secure at least ½*P* from a successful bargaining proposal, and this constitutes the limit beyond which O'Rourke and O'Reilly cannot expect her rationally to concede. Because the players in the game are not strategically equal, the side payments response does not prescribe an equal division of the surplus.

Thus, the side payments response prescribes a division (½*P*, ¼*P*, ¼*P)* in the Macquarrie-O'Rourke-O'Reilly game. Note that this agrees with Gauthier's conclusion as to what the payoffs should be – side payments lead us to the same result as the multi-stage bargaining upon which the cooperative infrastructures response relies. However, we have reached this result without encountering the path-dependence or arbitrariness problems raised by cooperative infrastructures. The strength of the side payments response is not that it delivers a different answer, but that it gets to the intuitively correct answer while avoiding the pitfalls in justification to which Gauthier's approach is vulnerable.[41]

[41] Compare Vanderschraaf, 2019: 322-323.

§8.3.3 Side payments in asymmetric contexts

We can get a clearer understanding of the side payments response by looking at how it performs in a couple of other asymmetric games. These are games, like the Macquarrie-O'Rourke-O'Reilly game, in which the players do not all have the same options, or the same productive capacities. This variability is likely to be a more realistic characterisation of the contractarian choice situation than uniformity.

[Game 6] is a majority-rule voting game in which there are four players (*A*, *B*, *C*, and *D*). *A* has two votes; each of the other players has one. The prize for any coalition commanding a majority of the votes is *P*, and the minimum winning coalitions are *AB*, *AC*, *AD*, and *BCD*. *A* can form a winning coalition with any of the other players; the other three players can between them muster the votes to exclude *A*. If we follow the logic of the side payments approach, any player who can make a disruptive offer must be paid off in order to prevent the offer being made. In the other games we have considered, the players with this power have either been symmetrically situated (as in the divide-the-dollar game) or have been necessary to *any* winning coalition (like Macquarrie). In this game, by contrast, *A* isn't essential, but might be considered to be in a stronger bargaining position than the other players, due to having more potential coalitions available.[42] Can *A* secure a better deal than the other players in this case?

I argue not. In **[Game 2]**, Macquarrie was able to secure a greater share of *P* than the other players. This was possible because of two (related) features of Macquarrie's strategic position, neither of which is shared by *A* in **[Game 6]**. First, Macquarrie could not be excluded from a winning coalition, whereas *A* gets nothing if the coalition *BCD* is formed. Second, Macquarrie could make an offer that could not be disrupted. This set a floor beneath which her payoff could not fall. In **[Game 6]**, any offer *A* makes is subject to disruption by some variant of the *BCD* coalition. And so, unlike Macquarrie, *A* always has an interest in making side payments to all of the

[42] Compare Riker and Ordeshook, 1973: 154-175. See also Laver, 1997: 138-140; Penrose, 1946; Banzhaf, 1965.

other players.

The strategically relevant features of a game which allow players to secure higher than equal payoffs are a player's necessity to the creation of a surplus, and vulnerability to disruptive offers. Since *A* is no more necessary to the creation of the surplus in **[Game 6]**, and no less vulnerable to disruptive offers, she is in exactly the same strategic position as the other players. The fact that *A* can be a part of a greater number of coalitions than the other players is not strategically relevant.[43] Thus, the only departure from strategic equality occurs in the case where some player is a necessary member of all possible winning coalitions. In other cases, any player who can be a necessary (non-dummy) member of *some* winning coalition can demand an equal share of the prize. (This will have important implications later, because it will mean many agents in the contractarian choice situation will have equal bargaining power.)

This result holds so long as the differences between the players are confined to voting weights. A further complication arises when we allow the size of the payoff to vary according to the membership of the winning coalition, as we did in **[Game 4]**. In **[Game 4]**, *all* players are essential to *some part* of the surplus, but the size of the additional surplus to which each player is essential varies. By contrast, in **[Game 5]** and **[Game 6]**, no player is essential – the whole surplus can be generated while excluding any one player from the winning coalition. As in the case of Macquarrie, being essential to the production of a particular portion of the surplus alters the strategic environment. A *bargaining floor* is created, beneath which a player cannot be forced to concede. These floors can be observed in the payoffs for the cooperative infrastructures approach to **[Game 4]**:

[43] At least in the ideal case. Transactions costs might make it relevant in less ideal circumstances, but I neglect that possibility here. See §5.5.2.

	A	*B*	*C*
Initial *AB*	$15(\frac{P}{12})$	$15(\frac{P}{12})$	$18(\frac{P}{12})$
Initial *AC*	$18(\frac{P}{12})$	$12(\frac{P}{12})$	$18(\frac{P}{12})$
Initial *BC*	$6(\frac{P}{12})$	$21(\frac{P}{12})$	$21(\frac{P}{12})$

Table 2[44]

A bargaining floor is the proportion of the surplus allocated to a player assuming that all other players are combined in a single coalition against her (call this the *complement coalition*). Given that both the excluded player and the complement coalition must agree in order to realise the relevant part of the surplus, half of the relevant part of the surplus is allocated to each. In **[Game 4]**, the whole of the surplus is 4*P*; setting a 'unit' of the surplus to $\frac{1}{12}P$, we have 48 units to distribute among the players. The bargaining floors for *A*, *B*, and *C* are then 6 units, 12 units, and 18 units respectively. That is, when *C* is excluded from the initial coalition, *A* and *B* can realise 12 units of the surplus without him. *AB* (the complement coalition) and *C* (the excluded player) are bargaining over the remaining 36 units of the surplus. Given that each is necessary, *C* is allocated half of these (18 units) and *AB* is allocated the other half (subsequently to be divided within that coalition in some way, *but not necessarily equally*). The other bargaining floors are calculated in a similar manner.

As no player can be forced to accept less than the value of his or her bargaining floor, any solution to this game must give at least 6 units to *A*, at least 12 units to *B*, and at least 18 units to *C*. The combined floors amount to 36 units, so the real

[44] This is a reproduction of Table 1, which set out the various payoffs associated with each choice of initial coalition in **[Game 4]**. For convenience, the denominators for the fractions of the surplus have now been set to 12. The bargaining floor for each player is highlighted.

problem a bargaining theory must solve is how to allocate the remaining 12 units.

Three options present themselves. First, there is a straightforward application of Gauthier's bargaining theory.[45] This has the effect of allocating the remaining units in proportion to the value of the floors: *A* receives 2 additional units, for a total of 8; *B* receives 4 additional units, for a total of 16; and *C* receives 6 additional units, for a total of 24. However, this is the solution both Gauthier and I have rejected because it fails to adequately capture the strategic options open to the players – this solution is the one that would allocate the whole of the surplus to Macquarrie in **[Game 2]**.[46]

The second possibility is to divide the remaining units equally between the three players. Each player would receive 4 additional units. The logic of the side payments approach might seem to recommend this option, since it appears to treat equally all players situated in strategically equal positions with respect to their capacity to make disruptive offers. However, I want to reject this 'locally egalitarian' option as well.

First, note that the locally egalitarian option is inconsistent with Gauthier's intuitions in the Macquarrie-O'Rourke-O'Reilly game. Macquarrie's bargaining floor is ½*P*. As inessential players, both O'Rourke and O'Reilly have bargaining floors of zero. The remainder of the surplus, once the sum of the three bargaining floors has been subtracted, is therefore also ½*P*. Applying the locally egalitarian option to this portion of the surplus yields O'Rourke and O'Reilly a payoff of $\frac{1}{6}P$ each, while Macquarrie would receive ½P (her bargaining floor) plus $\frac{1}{6}P$ (her equal share of the remaining surplus), for a total of ⅔*P*. By contrast, Gauthier's cooperative infrastructures approach allocated ½*P* to Macquarrie and ¼*P* to each of the assistants (as did side payments).

[45] This amounts to applying the Kalai-Smorodinsky solution, and equalising relative concessions, assuming that each player's maximum claim is that part of the surplus the other players cannot realise without him or her, and that the non-agreement point is 0. It is not possible to apply the Nash solution without settling on a mathematical representation for the players' bargaining power in **[Game 4]** (the exponents in the Nash solution), which is (part of) the point at issue. See §6.3, and the introductory remarks for this chapter.
[46] §8.2.2.

The locally egalitarian option takes bargaining floors as a *basis* for bargaining (an S_A).[47] That is, it treats bargaining floors as representing the payoff a player can secure in the absence of a bargain. It assumes that the complement coalitions have already been formed, and that they represent non-agreement points from which the parties are bargaining. It treats the part of the surplus beyond the sum of the bargaining floors as the only part of the surplus to which the players contribute equally in producing, thus treating the players as individually responsible for producing their own bargaining floors.

However, as we saw earlier, it is possible for players in **[Game 4]** to make offers which disrupt the formation of complement coalitions.[48] In refraining from making disruptive offers, players therefore contribute to the creation, not just of the surplus which is added when they join the grand coalition, but of the whole of the surplus.[49] Moreover, they do so equally. Insofar as each player is equally capable of making a disruptive offer, each can rationally demand equality in the *absolute* level he or she receives, and not simply equality in the share of that proportion of the surplus which exceeds the sum of the bargaining floors. That is to say, the players can demand that allocations be 'globally egalitarian'.

Strict global equality would require that each player be allocated 16 units (one third of the total available). Nonetheless, we also saw earlier that side payments to prevent disruptive offers do not always lead to an equal distribution of the available surplus.[50] In **[Game 2]**, being an essential player established a floor beneath which Macquarrie could not be pressed to rationally concede, because at any point below that floor she was capable of making an offer which could not be disrupted. A similar situation arises now in **[Game 4]**. Each of *A*, *B*, and *C* is essential to some part of the available surplus. An equal division of the surplus (16 units each) is above the bargaining floor for both *A* and *B*, but not for *C*.

---

[47] §6.1.
[48] §8.2.3.
[49] In joining the grand coalition, *A* increases the surplus available for distribution by *P*. However, in refraining from making disruptive offers, she increases the surplus by 4*P*.
[50] §8.2.3.

Suppose the initial proposal in the game is from *A* to *B*. The coalition *AB* can secure 12 units by itself, plus half of the surplus generated by adding *C* to the coalition (36 units). *A* and *B* would thus have 30 units (12 plus 18) to distribute between them. Suppose *A* initially proposes to divide these equally, such that each player receives 15 units. *C* can disrupt this offer by proposing an initial coalition *AC*.[51] *AC* can secure 24 units, plus half of the surplus generated by adding *B* (24 units), meaning that *A* and *C* have 36 units to divide. If *C* proposes that they split these equally, *A* will be receiving 18 units, and so prefers this offer to the original proposal. Importantly, *B* cannot disrupt this offer by *C*. Under this offer, *B* is receiving 12 units. Under the original offer, *A* and *B* were dividing 30 units. Any offer by *B* which allocates more than 18 units to *A* would leave *B* with fewer than 12 units, worse off than under *C*'s proposal. Thus, *C* cannot secure fewer than 18 units – his bargaining floor.[52]

A successful side payments allocation must respect both the concern with absolute or 'global' equality *and* the existence of bargaining floors. Rather than taking bargaining floors as a basis for bargaining, as the locally egalitarian solution does, we should understand the floors as a *constraint* on bargaining. How do we do this? First, give to each player an allocation equal to the value of his or her bargaining floor. Then, allocate the remaining surplus to the player with the lowest absolute level (in **[Game 4]**, *A*) until that player's absolute level reaches parity with the absolute level of the player with the next lowest absolute level. Continue to allocate any remaining surplus to each of these players equally, until they reach parity with the absolute level of the player with the next lowest absolute level. Repeat this process until one exhausts the surplus or one reaches a stage where all players are at the same absolute level; if a surplus still remains once all the players have reached the same absolute level, continue to allocate the surplus equally until it is exhausted. Call this the globally egalitarian or *prioritarian* option. In **[Game 4]**, the prioritarian option recommends a distribution of (15, 15, 18).

[51] Or *BC* – the argument proceeds in a similar way when *B* is substituted for *A*.
[52] And similarly, *mutatis mutandis*, for the other players and bargaining floors.

Absolute equality in the distribution of the surplus is rejected because it fails to account for the fact that players can secure their bargaining floors by making offers which cannot be disrupted. Of the three options for distributing the surplus which respect the existence of bargaining floors, the proportional option is rejected because it fails to fully account for the strategies available to the players. The locally egalitarian option is rejected because it fails to appreciate that, by refraining from disruptive offers, players contribute to the production of the whole of the coordination surplus.

The prioritarian option is accepted because it is consistent both with the equal claim players have on the whole of the surplus *and* with the existence of bargaining floors. It prevents disruptive offers, because it ensures that a player does as well as he or she can possibly do, in absolute terms, subject to the constraints that (i) no other player with the same capacity to make a disruptive offer does worse and (ii) no player receives less than the value of his or her bargaining floor. Given the equal capacity to make disruptive offers (that is, assuming we are talking about non-dummy players), I cannot rationally expect you will concede to a lower absolute level than me, except where allocating the surplus equally would give me an absolute level falling below my bargaining floor.

§8.4 Global rationality

The key move in the argument for the side payments response is the construction of 'global' rationality. Clearly, this is a revision of the standard understanding of rationality. On the standard account, it would be irrational to adopt the side payments approach, because it would commit one, for example, to 3-player coalitions in games where one could benefit by adopting a 2-player coalition and excluding the third player. On the standard account, it is always rational to make and accept disruptive offers.[53] By contrast, I claim it is (globally) rational for agents

[53] I am again assuming no time-preferences here (see n6). Making certain assumptions about preferences for present gains over future gains is one of the ways in which procedural models of

to avoid such offers in favour of side payments.

At this point, it might reasonably be objected that one can describe any bargaining solution whatsoever as 'rational' so long as one is free to define 'rationality' as one pleases. The success of the side payments response depends on a defence of global rationality.

My defence rests on three claims. First, the standard account of rationality is intuitively inadequate. It requires us to accept that, in games without cores, agents do nothing irrational in bargaining indefinitely. This is deeply implausible. An adequate conception of rationality must not endorse infinite cycling.

Second, there is nothing illicit about adjusting one's conception of rationality to respond to intuitively implausible implications.[54] This is how the money-pump problem leads us to endorse completeness and transitivity as necessary conditions of rationality.[55] A conception of rationality which allows incomplete preferences (for instance) forces us to say that an agent acts rationally when being money-pumped. This is intuitively implausible, and so our (pre-completeness) conception of rationality must be supplemented by a completeness condition. Since the standard conception has similarly implausible implications in games without cores, it must be supplemented in a manner which avoids these implications. The conception of global rationality I've described does this.

---

bargaining such as the alternating offers game impose sufficient structure to avoid infinite cycling. One might then ask, why exclude time-preferences from the characterisation of the coalition problem? First, it looks like some ways of building structure into models using time-preferences will amount to a form of exogenous determination. If so, the contractarian theory fails. Second, and more substantively, the structure imposed by time-preferences, however appropriate for a variety of bargaining situations, does not seem appropriate in the context of a contractarian choice situation. Agents in such a situation are idealised in respects which rule out time-preferences (considerations from §7.4 apply; see also the discussion in the Interface). That said, I wish to reject the implication that excluding time-preferences amounts to assuming a kind of (moralised) impartiality (Thoma, 2015: 3351). The presence of bargaining floors and dummy players in the side payments response are departures from impartiality sufficient for us to say that the response has not smuggled in a moral commitment to impartiality.

[54] Compare Broome, 1999: 413; Thoma, 2015: 3349-3350.

[55] §5.3, §5.4.

Third, I claim that the manner in which I have revised the standard conception of rationality to deal with the coalition problem is the best of the available solutions. Now, in discussing the money-pump problem earlier, I noted that there are several ways in which one might go about blocking the money-pump. I endorse the completeness and transitivity conditions, but there are others which would accomplish the same end.[56] The same is true of the coalition problem – the side payments solution is one way of blocking infinite cycling and exogenous determination, but it is not the only way. One simple way to endogenously block cycling would be to make it an axiom of rational choice that agents accept the first offer that is made and reject all subsequent offers. This does prevent cycling (and does so endogenously, because it is not an institutional constraint which blocks disruptive offers, but something internal to the agents' rationality). Still, as an alternative to the side payments response, this is obviously a less plausible extension of classical rationality. It commits agents to accepting highly disadvantageous proposals, for no particular reason. In linking prescribed divisions to strategic power, the side payments response has a plausibility that such simplistic alternatives lack.

Consider a more plausible alternative.[57] In the divide-the-dollar game, suppose we take the larger bargaining problem to be made up of two components. First, there is the question of which winning coalition to form. Second, there is the question of how to distribute the prize among members of that winning coalition. On the standard account of rationality, winning coalitions must be minimal; and 2-player coalitions, in which both parties are necessary members, divide a prize equally between the two parties. In the divide-the-dollar game, then, the standard account can answer the second question: Whichever minimum winning coalition is formed, each member of it should be allocated ½*P*. The problem for the standard account – the reason some extension is needed – is its inability to uniquely determine a winning coalition.

[56] §5.4.3.

[57] I thank David Miller, Sam Kiss, and Anthony Taylor for discussion of the issues raised in the next few paragraphs.

Suppose we extend the standard account as follows. In the divide-the-dollar game, each player flips a fair coin to determine to which of the other two players he or she will propose an equal division of the prize. In the event that this procedure leads to two players proposing the same division, the included players resolve to refuse all subsequent offers.[58] In the event that it does not, all coins are flipped again.[59] Call this the *mixed strategy* response. This procedure uniquely identifies a winning coalition, thus avoiding infinite cycling. On this extension of the standard account of rationality, resolutely refusing disruptive offers is a necessary condition of being rational, and so cycling is avoided endogenously. The mixed strategy response is thus a viable extension of the standard account of rationality.[60]

The mixed strategy response has some attractive features. The idea of using randomness to select one's action is a familiar feature of game theoretic analysis. The mixed strategy response also retains the standard account's commitment to minimum winning coalitions, where the side payments response endorses more inclusive coalitions (and hence lower payoffs for members of the winning coalition). In order to deal with scenarios more complicated than the divide-the-dollar game, some further specification of the procedure will be required. Nonetheless, the parameters of such specification should be reasonably clear. First, we identify all possible minimum winning coalitions. Second, we have each player use some randomisation device to select a winning coalition. Whichever coalition emerges from this random procedure is then implemented by the necessary players, who ignore attempts by excluded players to subvert the coalition with disruptive offers.

In the divide-the-dollar game, under the mixed strategy described above, two of the players will split P equally, while the remaining player will get nothing. Note, however, that the *expected* payoff for each player is $\frac{1}{3}P$ – the division proposed by

[58] Suppose, for example, that *A*'s coin selects *B* as a coalition partner for her, *B*'s selects *A*, and *C*'s selects *A*. Since *A* and *B* have both selected the coalition *AB*, this is the coalition they implement, resolving to refuse all subsequent disruptive offers from *C*. For further discussion of resolve in decision theoretic contexts, see §5.3.

[59] The procedure is indeterminate if the flips produce a cycle: *A*'s coin selects *B*, *B*'s selects *C*, and *C*'s selects *A* (or the reverse of this cycle).

[60] For an objection to extending the standard account of rationality in this way, based on the fact that it gives agents resolute commitment power, see Moehler (2018: 103-107).

the side payments response. That the two responses converge in this case is an artefact of the game's simplicity. In other games, the responses diverge. To properly compare the responses we must look at a different game.

Consider **[Game 6]** again. Recall that this is a majority-rule voting game in which there are four players (*A*, *B*, *C*, and *D*). *A* has two votes; each of the other players has one. The prize for any coalition commanding a majority of the votes is *P*, and the minimum winning coalitions are *AB*, *AC*, *AD*, and *BCD*. Because *A* has the option of three coalitions, rather than two, flipping a regular coin won't do. But suppose we stick with the spirit of the coin toss and say that *A* assigns each of her possible coalitions an equal probability (⅓). Suppose *B*, *C*, and *D* flip fair coins, as before. Given these assumptions, the expected payoff for *A* is ⅓*P*. The expected payoff for each of *B*, *C*, and *D* is $\frac{2}{9}P$.[61] By contrast, the side payments response gives each of the four players ¼*P*.[62] Compared to the side payments response, the mixed strategy response rewards *A* for (arguably) being in a stronger bargaining position relative to the other players. Insofar as we think *A* is in a stronger position, and think that payoffs should track bargaining power, this is a point in favour of the mixed strategy response.

The particular payoffs recommended by the mixed strategy response in this game are a function of the probabilities the players assign to each of their options. In the divide-the-dollar game, it seemed natural for each player to assign an equal probability to each option (that is, to flip fair coins). In illustrating the voting game, I have imported the equal probability assumption. But a mixed strategy need not consist of equal probabilities. The players could use any probabilities. Is it rational for them to use equal probabilities? More generally, is there a set of probabilities it is uniquely rational for them to use?

This question amounts to asking if there is a set of probabilities which it would be to

[61] Here, and below, I omit calculations to conserve space, but it should be fairly easy for a reader to confirm the various figures.
[62] All players are inessential, and so have bargaining floors of zero; all players can make disruptive offers, and so each is allocated an equal payoff (§8.3.3, §8.3.4).

a player's greatest advantage to choose. Once we formulate the problem in this way, we can reject the equal probability assumption as a general rule. In the voting game, suppose that *B*, *C*, and *D* each choose the coalition *BCD* with probability 1 (certainty). In that case, each of the three players would have an expected payoff of ⅓*P*, compared to $\frac{2}{9}P$ under equal probabilities. It is not always to the advantage of *B*, *C*, and *D* to choose among their options with equal probability. However, if we reject equal probabilities, what probability assignations should we be assuming?

⅓*P* is not the highest possible payoff available for *B*, *C*, or *D* under a mixed strategy. Each could achieve ½*P* in coordination with *A*. So, if *C* and *D* (say) are choosing *BCD* with probability 1, then *B* can achieve ½*P* by choosing *AB* with probability 1 (and similarly for *C* and *D*, *mutatis mutandis*).[63] It could thus be to *B*'s advantage to have a disposition to select probabilities which is conditional on the behaviour of *C* and *D*. However, what is rational for *B* will also be rational for the similarly situated *C* and *D*. If *B* has a disposition to choose *AB* which is conditional on *C* and *D* choosing *BCD*, then *C* and *D* will have similar conditional dispositions. But, if all three players have such a disposition, they will all end up choosing coalitions with *A*, and their expected payoffs will be $\frac{1}{6}P$.[64]

The problem with choosing *BCD* with certainty is that it creates incentives for the other players to defect. We can try to get around this by looking for a set of probabilities which do not create such incentives. A set with this property is called a *mixed strategy Nash equilibrium* (MSNE).[65] At a MSNE, no player can benefit by unilaterally varying the probabilities he or she assigns to his or her options. The MSNE for **[Game 6]** has *A* choosing among her options with equal probability, and

[63] It may appear that the expected payoff to *B* of choosing *AB* will depend on the probability with which *A* chooses *AB*. In particular, it may appear that the expected payoff to *B* will be just $\frac{1}{6}P$ if *A* chooses among *B*, *C*, and *D* with equal probability (⅓), and so that *B* will not benefit from switching to *AB* from *BCD*. But recall that the coalition selection procedure is reset whenever no viable coalition results from randomisation. In the present case, *C* and *D* are choosing *BCD* with probability 1, and *B* is choosing *AB* with probability 1. So long as *A* is choosing *AB* with *any* positive probability, *B* has an expected payoff of ½*P*, because all results other than *AB* will be discarded.

[64] The *B*-*C*-*D* subgame is a variation on the prisoner's dilemma – all players would benefit if they could coordinate on *BCD*, but each is better off defecting from *BCD* no matter what the other players do, and so they all end up worse off.

[65] Osborne, 2004: 99-119.

has *B*, *C*, and *D* choosing a coalition with *A* with probability of approximately 0.59, and *BCD* with probability of approximately 0.41. The expected payoff for *A* is approximately 37% of *P*, with the remaining players each expecting 21% of *P*.

I've now canvassed three possible sets of probabilities which could be used in a mixed strategy response in the voting game. I take it equal probabilities can be rejected. On a standard game theoretic account, the idea that *B*, *C*, and *D* could rationally coordinate on choosing *BCD* with certainty is untenable, and so a MSNE is the recommended solution. But we need to be careful about this conclusion. On a standard account, the reason coordination on *BCD* is untenable is that it is a dominated strategy – *whatever C* and *D* plan to do, *B* is better off choosing *AB*. But, because it is a response to the coalition problem, the mixed strategy response has already built into its agents a rational disposition not to defect when the standard account says they would (namely, when subject to disruptive offers). If we're suspending the rationality of accepting disruptive offers, why not also suspend the rationality of rejecting dominated strategies when they lead to suboptimal outcomes?[66]

One problem with coordination on *BCD* as an alternative to the MSNE is that giving a general operationalisation of the strategy is not straightforward. In the specific context of **[Game 6]**, the strategy could be stated as 'resolutely join the largest possible minimum winning coalition', but this will not achieve the desired result in games where players such as *B*, *C*, and *D* differ in their productive capacities. We would also have to consider what kinds of dispositions it would be rational for players such as *A* – who would be excluded from winning coalitions – to adopt in light of it being rational for other players to coordinate on *BCD*. For instance, it might be rational for *A* to have a disposition, conditional on being a player excluded by a *BCD* coalition, to be willing to resolutely join a coalition with one of the other players on terms disproportionately favourable to the other. So long as these terms are stacked sufficiently heavily against *A*, it may no longer be rational for *B*, *C*, and

[66] Binmore, 1994: 26-28, 173ff.

D to coordinate with certainty on *BCD*. Similarly, it is surely rational to have a disposition to accept disproportionately disadvantageous offers, on the condition that one has been identified as an excluded player by the operation of the randomisation device (and hence will otherwise be excluded entirely); this being the case, if one is an included player, it is rational to be disposed to take advantage of such dispositions instead of resolutely sticking with the coalition identified by the device. Once one is free to construct complex conditional dispositions for one's agents, it becomes hard to see where the process can stop – a version of the coalition problem will reappear in a different form, at the level of selecting conditional dispositions.

The mixed strategy response thus appears to be internally incoherent. Coordination on a MSNE is incoherent because it allows agents to act resolutely at one level (resisting disruptive offers), while simultaneously denying them that option at another level (resisting defection from coalitions when each doing so results in suboptimal payoffs). But, in the absence of coordination on a MSNE, there is no obvious way of preventing the regress of conditional dispositions from undoing the work that the ability to act resolutely was supposed to do in solving the coalition problem.

My conception of global rationality avoids these difficulties. It does not rely on exotic psychological devices such as resolve and conditional dispositions. Any (non-dummy) agent unhappy with the division proposed by the side payments response is free to make a disruptive offer. But, if she does so, other agents can respond in kind, and then the original agent will be worse off than she was under the division initially proposed by the side payments response. Global rationality consists in anticipating this and refraining from making or accepting disruptive offers when a division of a particular kind is proposed. It remains to ask, a division of what kind? The answer is given by the side payments response.

One of the upshots of this conception of global rationality is that the practical stability challenge posed in Chapter 3 can be answered. Practical instability results

from the possibility of rapidly shifting coalitions in a social contract theory. But the side payments response is predicated precisely on eliminating such instability – it is a solution to the coalition problem precisely because it argues for the rationality of a particular stable coalition which includes all non-dummy players and (allowing for adjustments based on bargaining floors) allocates to each an equal share of the social product.

§8.5 A basis of (strategic) equality

I have argued that the coalition problem can be solved, and that the side payments response is the best available solution. I now return to the question of bargaining power in the contractarian model.

On the side payments approach, bargaining power is a function of agents' strategic capacities. I have identified three capacities as relevant. First, the capacity to be a necessary member of some coalition which is effective in producing payoffs for its members (a non-dummy player). Second, the capacity to make disruptive offers. Third, the capacity to make offers which cannot be disrupted (to be an essential player). These capacities interact to determine a distribution of bargaining power. In order to have *any* bargaining power, an agent must have the first capacity. If an agent has the first capacity, she has the second to the extent that other agents lack the third.

Begin by examining the capacity to make disruptive offers. For my purpose, the key feature of this capacity is that it is what Rawls has called a 'range property'.[67] With the exceptions of dummy players and essential players, all agents in the contractarian model are equally possessed of the capacity to make disruptive offers; they do not possess the capacity in varying degrees. Range properties are important in liberal theory, because they can serve as a *basis of equality* – a justification for citizens having equal status despite varying over a range in many of the underlying

[67] Rawls, 1999a: 441-445.

dimensions which might qualify them for this status. In setting out the side payments approach, I have shown how agents with quite different abilities to contribute to production of a coordination surplus can nonetheless rationally demand an equal share of the surplus if they have the capacity to block the realisation of the surplus by making disruptive offers.[68]

Recall that the surplus of coordination in the contractarian model consists of degrees to which agents' conceptions of the good and justice are realised. Thus, in saying that the model allocates to all agents with the capacity to make disruptive offers an equal share of the surplus, we are saying that agents with vastly different levels of strength, talent, or productivity can nonetheless be entitled to demand equality in the protection and promotion of their concerns by the state. In §7.8, I argued that, assuming equal bargaining power and a society characterised by pluralism, what this will amount to in practice is a set of political institutions which guarantees basic rights, democratic participation, and substantively egalitarian distributions of material resources. Because the capacity to make disruptive offers is a range property, and because it is among the key capacities which determines bargaining power in the contractarian model, the side payments approach gives us an argument for assuming equal bargaining power. Thus, we have a contractarian basis of equality.

The question remaining is how much the two exceptions – dummy players and essential players – distort this egalitarianism. To recapitulate, dummy players are players who cannot be a necessary part of any winning coalition. Putting that another way, dummy players are incapable of affecting the degree to which any other agent's goals are realised – they fail to meet what I have termed the interaction condition.[69] As such, they have no bargaining power whatsoever in the contractarian model. They are not in the circumstances of justice with the other agents.

Earlier, I set aside the problem of agents who fail to meet the interaction condition,

[68] §8.3.3.
[69] §5.5.2. See also Vanderschraaf, 2019: 114-115.

because in order to defend the usefulness of the contractarian model it wasn't necessary to show that it applied to all agents. Here, however, a different worry arises. If, in the contractarian model, rights and entitlements are accorded only to agents with some degree of bargaining power, dummy players will be entirely excluded. To what extent is this exclusion a problem for the model as a *liberal* theory?[70]

There are three things to say here. The first response is Hume's: Justice is a virtue governing interaction where there is the possibility of mutual advantage. We may well have moral reasons to give due regard to the interests of dummy players, but they are beyond the scope of justice, liberal or otherwise.[71]

Second, recall how weak the demands of the interaction condition are. It is sufficient, to avoid being a dummy player, that one can affect the realisation of some other agent's goals in *some* degree.[72] Importantly, the baseline against which we measure is *not* non-interaction.[73] Rather the baseline is Hobbesian, and so if you refrain from assaulting me, you thereby contribute to the realisation of my goals. No person who is capable of any kind of assault on me can be a dummy player (and it seems trivial to assume there will be at least some agents who, like me, have preferences for minimising incidences of assault). Now, many agents are capable of some kind of assault, but there may also be agents whose abilities do not extend so far. Happily, being capable of assault is not a *necessary* condition of mutually advantageous interaction. Any person who can raise an alarm if you try to assault me while I sleep can aid in the realisation of my goals. Considerations such as these lead me to think that very few agents will end up being dummy players.[74]

---

[70] See Moehler, 2018: 86-87. Vanderschraaf (2019: 280-292) provides a detailed discussion of one variation on this problem, which he terms the Vulnerability Objection.
[71] §5.5.2.
[72] Compare Vanderschraaf, 2019: 283-284.
[73] §7.5.
[74] Note, at this point I am still assuming that agents are rational (and thus can make and keep agreements). Certain kinds of entities – young children, non-human animals, people with certain brain injuries or cognitive disabilities (compare Moehler, 2018: 168, 170-172) – are still excluded from consideration. With respect to such entities, the first or third response is required. Again, see Vanderschraaf (2019: 280-304).

Third, recall what is at stake between the agents in the contractarian model. They are bargaining over (among other things) the extent to which they can realise their conceptions of justice. It is one thing to worry that dummy players will not be entitled to the equal rights other citizens can demand. But that's not what's really at issue. Rather, what dummy players are demanding is the implementation (to some degree) of their conceptions of justice. This, we should have no difficulty denying them. So long as we are liberals, and they are agents of the right kind, their *interests* will be protected by the implementation of *our* conception of justice, without the need to concede to them any share of the capacity to determine which conception of justice will be implemented.[75] That is to say, in the pursuit of justice, greed is no vice.[76] As such, exclusion is no problem.

For these reasons, it does not seem to me that dummy players pose much difficulty for my argument. What about essential players? Recall that a player is essential if any part of the coordination surplus cannot be generated without her contribution. A player is inessential if the role she plays in producing the surplus could be filled equally well by someone else.[77] Essential players have (non-zero) bargaining floors, and these can lead to inequalities in the distribution of payoffs.

There are two questions to consider here. First, how likely is it that there will be any essential players? Second, given that there are some essential players, will this lead to inequality? On the one hand, we might think that essential players are rare. For how many of us is it the case that our contribution to the social product is absolutely irreplaceable? For some of us, sure. But perhaps not many – even without Leibniz, we would still have Newton. This thought is too quick, however. In discussing

[75] Moehler, 2018: 170-171. Obviously, the antecedent is important here. See the discussion of *modus vivendi* in the Introduction. For some (moral) concerns about a response of this form, see Vanderschraaf (2019: 285-287), but note that these concerns will not apply when the distribuendum is 'capacity to determine the content of political institutions' rather than 'resources' or 'goods'.
[76] Compare Gaus, 2018: 126-127.
[77] Thus, inessential players are not dummy players. Inessential players can be substituted for one another; dummy players cannot substitute for an inessential player. Compare Vanderschraaf's discussion of (negative and positive) mutual expectations, the contribution requirement, and the Vulnerability Objection (2019: 276-292).

dummy players, I noted that an agent's contribution to the coordination surplus includes such simple things as refraining from assault. This is a contribution no one else can make. In that sense, at least, perhaps most of us are essential players.

If many of us are essential players, the next question is the extent to which this leads to inequality in the contractarian model. Recall how the prioritarian version of the side payments response works.[78] First, we allocate to each player her bargaining floor. We subsequently allocate any additional surplus to the least advantaged players until we either exhaust the surplus or we reach a position where all players are at an equal level and above their bargaining floors. This is to say, bargaining floors *can* lead to inequalities in the distribution of the surplus, but do not when the size of the surplus is sufficiently great relative to the sum of the bargaining floors. Inequality in the contractarian model is therefore a product of there not being very great gains from coordination or of many agents having high bargaining floors, and thus being essential to large proportions of the surplus. We have reason to doubt that either of these conditions will obtain. I have already discussed the likely extent of gains from coordination.[79] And, while I was happy to concede that many of us make *some* essential contribution, simply in virtue of refraining from attacking one another, this contribution does not seem to give rise to a particularly high bargaining floor. My refraining from assaulting you may be of great importance to you, but it is a small fraction of the social product. In the event that bargaining floors are not high, their effects are swamped by the prioritarian distribution rule.

It seems to me, then, that the prospect of essential players does not significantly undermine the extent to which the side payments response gives rise to substantively egalitarian distributions of bargaining power. Moreover, we have reason to doubt that all such inequalities are incompatible with liberalism. In Rawlsian terms, the features of individual agents which give rise to bargaining floors in the contractarian account would be due either to natural primary goods (advantages which lead to greater productivity, and hence a high bargaining floor), or to the exercise of basic

[78] §8.3.3.
[79] §7.5.

liberties (on the Rawlsian account, people can't be forced to engage in cooperation in order to increase the social product, they must be given an incentive to do so). Inequalities arising from either of these sources are consistent with Rawlsian liberalism.

In the preceding discussion, I have not considered what rational bargaining agents would do in circumstances where the overall coordination surplus varies according to the level allocated to certain players – the classic case to which the difference principle is applied – but I see no reason to think that the contractarian prescription will diverge from the allocations supported by the difference principle, for precisely the reasoning that Rawls gives. Mutually disinterested rational agents, in circumstances where the least advantaged agent has a veto over the distribution of the surplus, will arrive at something like the difference principle. The crucial component of either argument is the veto – once that is established, prudential rationality does the rest. Rawls supports the veto by appeal to what he sees as reasonable moral constraints on the circumstances of agreement; my argument has been that, in a wide variety of cases (although not in all cases), the veto can be defended simply by appeal to prudential rationality, exercised in circumstances subject to the coalition problem.

Conclusion

In Part II, I have argued that contractarianism – a social contract theory relying not on moral reasons, but solely on prudential reasons – endorses robustly liberal and egalitarian political institutions. Such institutions are the uniquely rational choice for agents who are relevantly idealised versions of actual citizens, more rational than real people, but no less diverse in their tastes, ambitions, and moral commitments.

Of course, the neat thing about social contract theories is that you can get anything you want out of them, so long as you finesse their inputs in just the right way. The strength of my conclusion depends on the plausibility of the claims I have made about the agents, their circumstances, and the nature of rational bargaining. Even if my argument is internally valid, to be convinced of my conclusion you must think that the characterisations I have given to the various elements of the contractarian model are relevantly congruent with real people and their circumstances, and that the idealisations I have made are justified without the need for moral appeals. But I think they are, and have argued as much in Chapters 5-8.

Why should we *care* that contractarianism endorses liberal political institutions? I began, some hundred thousand words ago, by asking how one can best pursue one's conception of justice when the people upon whom its realisation depends cannot be persuaded that it is correct. In Part I, I argued that, given pervasive and persistent moral disagreement, contractarianism is essentially inescapable in *political* theory. If we have a duty to advance justice as best we can, cannot persuade other people of the truth of our conception of justice, and lack the power to simply impose it, we must be committed to entering into social arrangements with others on the basis only of prudential reasons. It thus matters very much how far prudential reasons and contractarianism can take us.

I present, then, a *modus vivendi* liberalism. It is true that this account of liberalism

cannot guarantee the inclusion of *all* people, or that there will *never* be tyrants or extremists so powerful that the coalition conditions upon which my egalitarianism rests will not be met. But no account which takes seriously the need to give effective reasons in politics could make such a guarantee. So long as there remains profound disagreement about justice, this is the best we can do. It need not be all to which we aspire. I have not said that a consensus on justice is beyond us. For here and now, though, we do not need one.

Bibliography

Abizadeh, Arash. 2011. "Hobbes on the Causes of War: A Disagreement Theory" in *American Political Science Review* 105:2, pp. 298-315.

Ackerman, Bruce. 1980. *Social Justice in the Liberal State*. New Haven, Connecticut: Yale University Press.

Aldred, Jonathan. 2007. "Intransitivity and Vague Preferences" in *The Journal of Ethics* 11:4, pp. 377-403.

Almeida, Michael J. (ed.). 2000. *Imperceptible Harms and Benefits*. Dordrecht: Kluwer Academic Publishers.

Altson, William P. 1976. "Two Types of Foundationalism" in *The Journal of Philosophy* 73:7, pp. 165-185.

Anand, Paul. 1987. "Are the Preference Axioms Really Rational?" in *Theory and Decision* 23:2, pp. 189-214.

Anand, Paul. 1993. "The Philosophy of Intransitive Preference" in *The Economic Journal* 103:417, pp. 337-346.

Aristotle. [1981]. *The Politics*. Translated by T. A. Sinclair, revised and re-presented by Trevor J. Saunders. London: Penguin Books.

Arneson, Richard J. 1990a. "Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare" in *Philosophy & Public Affairs* 19:2, pp. 158-194.

Arneson, Richard J. 1990b. "Primary Goods Reconsidered" in *Noûs* 24:3, pp. 429-454.

Baderin, Alice. 2014. "Two forms of realism in political theory" in *European Journal of Political Theory* 13:2, pp. 132-153.

Banzhaf, John F. (III). 1965. "Weighted Voting Doesn't Work: A Mathematical Analysis" in *Rutgers Law Review* 19:2, pp. 317-343.

Bar-Hillel, Maya and Avishai Margalit. 1988. "How Vicious Are Cycles of Intransitive Choice?" in *Theory and Decision* 24:2, pp. 119-145.

Barry, Brian. 1989. *A Treatise on Social Justice Volume I: Theories of Justice*. London: Harvester-Wheatsheaf.

Barry, Brian. 1995. *A Treatise on Social Justice Volume II: Justice as Impartiality*.

Oxford: Oxford University Press.

Billingham, Paul and Anthony Taylor. MS. "Idealization, Rationale, and Content: A Framework for Analyzing Public Reason Theories" (draft paper provided by the authors 14 August 2019).

Binmore, Ken. 1993. "Bargaining and morality" in *Rationality, Justice and the Social Contract: Themes from* Morals by Agreement, eds. David Gauthier and Robert Sugden. New York: Harvester Wheatsheaf.

Binmore, Ken. 1994. *Game Theory and the Social Contract Volume I: Playing Fair*. Cambridge, Massachusetts: The MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract II: Just Playing*. Cambridge, Massachusetts: The MIT Press.

Binmore, Ken and Alex Voorhoeve. 2003. "Defending Transitivity Against Zeno's Paradox" in *Philosophy & Public Affairs* 31:3, pp. 272-279.

Bratman, Michael E. 2013. "The Interplay of Intention and Reason" in *Ethics* 123:4, pp. 657-672.

Broome, John. 1978. "Choice and Value in Economics" in *Oxford Economic Papers* (New Series) 30:3, pp. 313-333.

Broome, John. 1995. *Weighing Goods: Equality, Uncertainty and Time*. Oxford: Basil Blackwell.

Broome, John. 1999. "Normative Requirements" in *Ratio* (New Series) 12:4, pp. 398-419.

Broome, John. 2005a. "Does Rationality Give Us Reasons?" in *Philosophical Issues* 15 (Normativity), pp. 321-337.

Broome, John. 2005b. "Have We Reason To Do As Rationality Requires? – A Comment on Raz" in *Journal of Ethics & Social Philosophy* 1:1; https://doi.org/10.26556/jesp.v1i1.13 (Accessed 04 August 2018.)

Broome, John. 2013. *Rationality Through Reasoning*. Chichester, West Sussex: Wiley-Blackwell.

Buchanan, James. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: The University of Chicago Press.

Carroll, Ian J. 2006. *On Scepticism, Neutrality, and the Social Contract*. MPhil Politics (Political Theory), University of Oxford.

Chambers, Simone. 2010. "Theories of Political Justification" in *Philosophy Compass* 5:11, pp. 893-903.

Chang, Ruth. 2002. "The Possibility of Parity" in *Ethics* 112:4, pp. 659-688.

Chappell, Sophie Grace. 2013. "Bernard Williams" in *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/spr2015/entries/williams-bernard/ (Accessed 04 August 2018.)

Christiano, Thomas. 1996. *The Rule of the Many: Fundamental Issues in Democratic Theory*. Boulder, Colorado: Westview Press.

Clark, William Roberts, Matt Golder, and Sona Nadenichek Golder. 2013. *Principles of Comparative Politics* (2nd Edition). Los Angeles: Sage/CQ Press.

Cohen, G. A. 2008. *Rescuing Justice and Equality*. Cambridge, Massachusetts: Harvard University Press.

Coram, Alex Talbot. 2001. *State, Anarchy and Collective Decisions*. Basingstoke, Hampshire: Palgrave.

Crowder, George. 2007. "Two Concepts of Liberal Pluralism" in *Political Theory* 35:2, pp. 121-146.

Cubitt, Robin P. and Robert Sugden. 2001. "On Money Pumps" in *Games and Economic Behavior* 37:1, pp. 121-160.

Cudd, Ann and Seena Eftekhari. 2017. "Contractarianism" in *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/sum2018/entries/contractarianism/ (Accessed 04 August 2018.)

D'Agostino, Fred. 1996. *Free Public Reason: Making It Up As We Go*. Oxford: Oxford University Press.

D'Agostino, Fred, Gerald Gaus, and John Thrasher. 2017. "Contemporary Approaches to the Social Contract" in *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/win2017/entries/contractarianism-contemporary/ (Accessed 04 August 2018.)

Dahl, Robert A. 2004. "What have we learned? Responses by Robert A. Dahl, Truman F. Bewley, Susanne Hoeber Rudolph, and John Mearsheimer" in *Problems and Methods in the Study of Politics*, eds. Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud. Cambridge: Cambridge University Press.

Dahl, Robert A. and Bruce Stinebrickner. 2003. *Modern Political Analysis* (6th Edition). Upper Saddle River, New Jersey: Prentice Hall.

Daniels, Norman (ed.). 1989. *Reading Rawls: Critical Studies on Rawls' A* Theory of Justice. Stanford, California: Stanford University Press.

Danielson, Peter. 1991. "The Lockean Proviso" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Darwall, Stephen (ed.). 2003a. *Contractarianism/Contractualism*. Oxford: Blackwell Publishing.

Darwall, Stephen. 2003b. "Introduction" in *Contractarianism/Contractualism*, ed. Stephen Darwall. Oxford: Blackwell Publishing.

Dreben, Burton. 2003. "On Rawls and Political Liberalism" in *The Cambridge Companion to Rawls*, ed. Samuel Freeman. Cambridge: Cambridge University Press.

Dworkin, Ronald. 1981a. "What Is Equality? Part 1: Equality of Welfare" in *Philosophy & Public Affairs* 10:3, pp. 185-246.

Dworkin, Ronald. 1981b. "What Is Equality? Part 2: Equality of Resources" in *Philosophy & Public Affairs* 10:4, pp. 283-345.

Dworkin, Ronald. 1985. *A Matter of Principle*. Oxford: Clarendon Press.

Dworkin, Ronald. 1989. "The Original Position" in *Reading Rawls: Critical Studies on Rawls' A* Theory of Justice, ed. Norman Daniels. Stanford, California: Stanford University Press.

Dworkin, Ronald. 1996. "Objectivity and Truth: You'd Better Believe It" in *Philosophy & Public Affairs* 25:2, pp. 87-139.

Eccleshall, Robert. 1994. "Liberalism" in *Political Ideologies: An Introduction* (2nd Edition), eds. Robert Eccleshall, Vincent Geoghegan, Richard Jay, Michael Kenny, Iain MacKenzie, and Rick Wilford. London: Routledge.

Eccleshall, Robert, Vincent Geoghegan, Richard Jay, Michael Kenny, Iain MacKenzie, and Rick Wilford (eds). 1994. *Political Ideologies: An Introduction* (2nd Edition). London: Routledge.

Elster, Jon. 1988. "The Nature and Scope of Rational-Choice Explanation" in *Science in Reflection* (The Israeli Colloquium: Studies in History, Philosophy, and Sociology of Science, Volume 3), ed. Edna Ullmann-Margalit.

Enoch, David. 2005. "Why Idealize?" in *Ethics* 115:4, pp. 759-787.

Enoch, David. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Enoch, David. 2015. "Against Public Reason" in *Oxford Studies in Political Philosophy* (Volume 1), eds. David Sobel, Peter Vallentyne, and Steven Wall. Oxford: Oxford University Press.

Farrelly, Colin. 2007. "Justice in Ideal Theory: A Refutation" in *Political Studies* 55:4, pp. 844-864.

Finkelstein, Claire. 2013. "Pragmatic Rationality and Risk" in *Ethics* 123:4, pp. 673-699.

Foot, Philippa. 2002. *Virtues and Vices and Other Essays in Moral Philosophy*. Oxford: Clarendon Press.

Freeden, Michael. 1996. *Ideologies and Political Theory: A Conceptual Approach*. Oxford: Clarendon Press.

Freeden, Michael. 2005. "What Should the 'Political' in Political Theory Explore?" in *The Journal of Political Philosophy* 13:2, pp. 113-134.

Freeman, Samuel. 2003. *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press.

Gaertner, Wulf and Marlies Klemisch-Ahlert. 1991. "Gauthier's approach to distributive justice and other bargaining solutions" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Galston, William. 1999. "Value Pluralism and Liberal Political Theory" in *The American Political Science Review* 93:4, pp. 769-778.

Gaus, Gerald F. 2000. *Political Concepts and Political Theories*. Boulder, Colorado: Westview Press.

Gaus, Gerald. 2003. *Contemporary Theories of Liberalism: Public Reason as a Post-Enlightenment Project*. London: SAGE Publications.

Gaus, Gerald. 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press.

Gaus, Gerald. 2014. "The good, the bad, and the ugly: three agent-type challenges to The Order of Public Reason" in *Philosophical Studies* 170:3, pp. 563-577.

Gaus, Gerald. 2018. "Self-organizing moral systems: Beyond social contract theory" in *Politics, Philosophy & Economics* 17:2, pp. 119-147.

Gaus, Gerald, Shane D. Courtland, and David Schmidtz. 2018. "Liberalism" in *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/spr2018/entries/liberalism/ (Accessed 04 August 2018.)

Gauthier, David. 1974. "Justice and Natural Endowment: Toward a Critique of Rawls' Ideological Framework" in *Social Theory and Practice* 3:1, pp. 3-26.

Gauthier, David. 1986. *Morals By Agreement*. Oxford: Oxford University Press.

Gauthier, David. 1988. "Moral Artifice" in *Canadian Journal of Philosophy* 18:2, pp. 385-418.

Gauthier, David. 1991. "Why contractarianism?" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Gauthier, David. 1993. "Uniting separate persons" in *Rationality, Justice and the Social Contract: Themes from* Morals by Agreement, eds. David Gauthier and Robert Sugden. New York: Harvester Wheatsheaf.

Gauthier, David. 1994. "Assure and Threaten" in *Ethics* 104:4, pp. 690-721.

Gauthier, David. 1995. "Public Reason" in *Social Philosophy and Policy* 12:1, pp. 19-42.

Gauthier, David. 1997. "Political Contractarianism" in *The Journal of Political Philosophy* 5:2, pp. 132-148.

Gauthier, David. 2013. "Twenty-Five On" in *Ethics* 123:4, pp. 601-624.

Gauthier, David and Robert Sugden (eds.). 1993. *Rationality, Justice and the Social Contract: Themes from* Morals by Agreement. New York: Harvester Wheatsheaf.

Geuss, Raymond. 2008. *Philosophy and Real Politics*. Princeton, New Jersey: Princeton University Press.

Gray, John. 2000. *Two Faces of Liberalism*. New York: The New Press.

Gustafsson, Johan E. and Nicolas Espinoza. 2010. "Conflicting Reasons in the Small-Improvement Argument" in *The Philosophical Quarterly* 60:241, pp. 754-763.

Hampton, Jean. 1986. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.

Hampton, Jean. 1991. "Two faces of contractarian thought" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Hampton, Jean. 1997. *Political Philosophy*. Boulder, Colorado: Westview Press.

Hansson, Sven Ove. 1989. "A New Semantical Approach to the Logic of Preference" in *Erkenntnis* 31:1, pp. 1-42.

Hansson, Sven Ove. 1993. "Money-Pumps, Self-Torturers and the Demons of Real Life" in *Australasian Journal of Philosophy* 71:4, pp. 476-485.

Hansson, Sven Ove. 2001. *The Structure of Values and Norms*. Cambridge: Cambridge University Press.

Hansson, Sven Ove and Till Grüne-Yanoff. 2017. "Preferences" in *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/sum2018/entries/preferences/ (Accessed 04 August 2018.)

Hargreaves Heap, Shaun P. and Yanis Varoufakis. 1995. *Game Theory: A Critical Introduction*. London: Routledge.

Hargreaves Heap, Shaun P., Martin Hollis, Bruce Lyons, Robert Sugden, and Albert Weale. 1992. *The Theory of Choice: A Critical Guide*. Oxford: Blackwell Publishers.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

Hayek, F. A. 1945. "The Use of Knowledge in Society" in *The American Economic Review* 35:4, pp. 519-530.

Heath, Joseph. 2015. "Methodological Individualism" in *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/entries/methodological-individualism/ (Accessed 16 October 2018.)

Held, David. 2006. *Models of Democracy* (3rd Edition). Cambridge: Polity Press.

Heywood, Andrew. 2003. *Political Ideologies: An Introduction* (3rd Edition). Basingstoke, Hampshire: Palgrave Macmillan.

Hobbes, Thomas. 1651 [1996]. *Leviathan* (Revised Student Edition). Edited by Richard Tuck. Cambridge: Cambridge University Press.

Horton, John. 2010. "Realism, liberal moralism and a political theory of modus vivendi" in *European Journal of Political Theory* 9:4, pp. 431-448.

Hubin, D. Clayton. 1979. "The Scope of Justice" in *Philosophy & Public Affairs* 9:1, pp. 3-24.

Hubin, Donald C. and Mark B. Lambeth. 1991. "Providing for rights" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Hume, David. 1739 [1984]. *A Treatise of Human Nature*. London: Penguin Books.

Hume, David. 1777 [1948]. "Of the Original Contract" in *Hume's Moral and Political Philosophy*, ed. Henry D. Aiken. New York: Hafner Press.

Hume, David. [1998]. *An Enquiry Concerning the Principles of Morals*. Oxford: Oxford University Press.

Hurka, Thomas. 1993. *Perfectionism*. Oxford: Oxford University Press.

Imai, Haruo. 1983. "Individual Monotonicity and Lexicographic Maximin Solution" in *Econometrica* 51:2, pp. 389-401.

Jaggar, Alison M. 1983. *Feminist Politics and Human Nature*. Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Kalai, Ehud. 1977. "Proportional Solutions to Bargaining Situations: Interpersonal Utility Comparisons" in *Econometrica* 45:7, pp. 1623-1630.

Kalai, Ehud and Meir Smorodinsky. 1975. "Other Solutions to Nash's Bargaining Problem" in *Econometrica* 43:3, pp. 513-518.

Kavka, Gregory S. 1983. "The Toxin Puzzle" in *Analysis* 43:1, pp. 33-36.

Kavka, Gregory S. 1986. *Hobbesian Moral and Political Theory*. Princeton, New Jersey: Princeton University Press.

Kiss, Sam. 2016. *Analysing and Evaluating Political Realism*. DPhil Politics, University of Oxford.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kramer, Matthew. 1999. *In Defense of Legal Positivism: Law Without Trimmings*.

Oxford: Oxford University Press.

Kraus, Jody S. 1993. *The Limits of Hobbesian Contractarianism*. Cambridge: Cambridge University Press.

Kymlicka, Will. 1989. "Liberal Individualism and Liberal Neutrality" in *Ethics* 99:4, pp. 883-905

Larmore, Charles E. 1987. *Patterns of Moral Complexity*. Cambridge: Cambridge University Press.

Larmore, Charles. 1990. "Political Liberalism" in *Political Theory* 18:3, pp. 339-360.

Larmore, Charles. 1996. *The Morals of Modernity*. Cambridge: Cambridge University Press.

Larmore, Charles. 2003. "Public Reason" in *The Cambridge Companion to Rawls*, ed. Samuel Freeman. Cambridge: Cambridge University Press.

Larmore, Charles. 2013. "What Is Political Philosophy?" in *Journal of Moral Philosophy* 10:3, pp. 276-306.

Laver, Michael. 1997. *Private Desires, Political Action: An Invitation to the Politics of Rational Choice*. London: SAGE Publications.

Lenman, James. 2009. "Reasons for Action: Justification vs. Explanation" in *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/spr2016/entries/reasons-just-vs-expl/ (Accessed 04 August 2018.)

Leopold, David and Marc Stears (eds.). 2008. *Political Theory: Methods and Approaches*. Oxford: Oxford University Press.

Leys, Wayne A. R. 1965. "Was Plato Non-Political?" in *Ethics* 75:4, pp. 272-276.

List, Christian. 2003. "Are interpersonal comparisons of utility indeterminate?" in *Erkenntnis* 58:2, pp. 229-260.

Locke, John. [2005]. "The Second Treatise of Government" in *The Selected Political Writings of John Locke*, ed. Paul E. Sigmund. New York: W. W. Norton & Company.

Macedo, Stephen. 1991. *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism*. Oxford: Clarendon Press.

MacIntosh, Duncan. 2013. "Assuring, Threatening, a Fully Maximising Theory of

Practical Rationality, and the Practical Duties of Agents" in *Ethics* 123:4, pp. 625-656.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth, Middlesex: Penguin Books.

Margalit, Avishai. 2009. *On Compromise and Rotten Compromises*. Princeton, New Jersey: Princeton University Press.

Matravers, Matt. 2000. *Justice and Punishment: The Rationale of Coercion*. Oxford: Oxford University Press.

May, Kenneth O. 1952. "A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision" in *Econometrica* 20:4, pp. 680-684.

McCabe, David. 2010. *Modus Vivendi Liberalism: Theory and Practice*. Cambridge: Cambridge University Press.

McClennen, Edward F. 1990. *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.

McDermott, Daniel. 2008. "Analytical political philosophy" in *Political Theory: Methods and Approaches*, eds. David Leopold and Marc Stears. Oxford: Oxford University Press.

McKelvey, Richard D. 1979. "General Conditions for Global Intransitivities in Formal Voting Models" in *Econometrica* 47:5, pp. 1085-1112.

Mill, John Stuart. 1859 [1974]. *On Liberty*. London: Penguin Books.

Moehler, Michael. 2018. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press.

Morris, Christopher W. 1991. "Moral standing and rational-choice contractarianism" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Morris, Christopher W. 2013. "Introduction" (to Symposium on David Gauthier's Morals By Agreement) in *Ethics* 123:4, pp. 595-600.

Moulin, Hervé. 1984. "Implementing the Kalai-Smorodinsky bargaining solution" in *Journal of Economic Theory* 33:1, pp. 32-45.

Muldoon, Ryan. 2016. *Social Contract Theory for a Diverse World: Beyond Tolerance*. London: Routledge. (Electronic Legal Deposit copy, accessed 23 September 2019.)

Narveson, Jan. 1988. *The Libertarian Idea*. Philadelphia: Temple University Press.

Narveson, Jan. 1991. "Gauthier on distributive justice and the natural baseline" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Nash, John. 1950. "The Bargaining Problem" in *Econometrica* 18:2, pp. 155-162.

Nash, John. 1953. "Two-Person Cooperative Games" in *Econometrica* 21:1, pp. 128-140.

Neal, Patrick. 1997. *Liberalism and Its Discontents*. Basingstoke, Hampshire: Macmillan Press.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell Publishing.

Oakeshott, Michael. 1991. *Rationalism in politics and other essays* (New and expanded edition). Indianapolis, Indiana: Liberty Fund.

Osborne, Martin J. 2004. *An Introduction to Game Theory*. Oxford: Oxford University Press.

Osborne, Martin J. and Ariel Rubinstein. 1990. *Bargaining and Markets*. San Diego, California: Academic Press, Inc.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Penrose, L. S. 1946. "The Elementary Statistics of Majority Voting" in *Journal of the Royal Statistical Society* 109:1, pp. 53-57.

Peterson, Martin. 2009. *An Introduction to Decision Theory*. Cambridge: Cambridge University Press.

Philips, Michael. 1989. "Must Rational Preferences Be Transitive?" in *The Philosophical Quarterly* 39:157, pp. 477-483.

Plato. [1993]. *Republic*. Translated by Robin Waterfield. Oxford: Oxford University Press.

Pollock, John L. 2006. *Thinking About Acting: Logical Foundations for Rational Decision Making*. Oxford: Oxford University Press.

Quinn, Warren S. 1990. "The Puzzle of the Self-Torturer" in *Philosophical Studies* 59:1, pp. 79-90.

Quong, Jonathan. 2011. *Liberalism Without Perfection*. Oxford: Oxford University

Press.

Quong, Jonathan. 2014. "What is the point of public reason?" in *Philosophical Studies* 170:3, pp. 545-553.

Quong, Jonathan. 2017. "Public Reason" in *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/archives/spr2018/entries/public-reason/ (Accessed 23 September 2019.)

Rabinowicz, Wlodek. 2000. "Money Pump With Foresight" in *Imperceptible Harms and Benefits*, ed. Michael J. Almeida. Dordrecht: Kluwer Academic Publishers.

Rachels, Stuart. 1998. "Counterexamples to the Transitivity of *Better Than*" in *Australasian Journal of Philosophy* 76:1, pp. 71-83.

Rawls, John. 1985. "Justice as Fairness: Political not Metaphysical" in *Philosophy & Public Affairs* 14:3, pp. 223-251.

Rawls, John. 1997. "The Idea of Public Reason Revisited" in *The University of Chicago Law Review* 64:3, pp. 765-807.

Rawls, John. 1999a. *A Theory of Justice* (Revised Edition). Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Rawls, John. 1999b. *The Law of Peoples with "The Idea of Public Reason Revisited"*. Cambridge, Massachusetts: Harvard University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*. Edited by Erin Kelly. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Rawls, John. 2005. *Political Liberalism* (Expanded Edition). New York: Columbia University Press.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Raz, Joseph. 1999. *Practical Reason and Norms*. Oxford: Oxford University Press.

Raz, Joseph. 2011. *From Normativity to Responsibility*. Oxford: Oxford University Press.

Rehfeld, Andrew. 2010. "Offensive Political Theory" in *Perspectives on Politics* 8:2, pp. 465-486.

Riker, William H. 1982. *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. Long Grove, Illinois:

Waveland Press, Inc.

Riker, William H. and Peter C. Ordeshook. 1973. *An Introduction to Positive Political Theory*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Ripstein, Arthur. 1987. "Foundationalism in Political Theory" in *Philosophy & Public Affairs* 16:2, pp. 115-137.

Rossi, Enzo. 2010. "Modus Vivendi, Consensus, and (Realist) Liberal Legitimacy" in *Public Reason* 2:2, pp. 21-39.

Rousseau, Jean-Jacques. 1755 [1984]. *A Discourse on Inequality*. Translated by Maurice Cranston. London: Penguin Books.

Rousseau, Jean-Jacques. 1762 [1968]. *The Social Contract*. Translated by Maurice Cranston. London: Penguin Books.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model" in *Econometrica* 50:1, pp. 97-109.

Ryan, Alan. 2012. *The Making of Modern Liberalism*. Princeton, New Jersey: Princeton University Press.

Sartre, Jean-Paul. 1948. *Existentialism and Humanism*. Translated by Philip Mairet. London: Methuen & Co. Ltd.

Saunders, Ben. 2010. "Democracy, Political Equality, and Majority Rule" in *Ethics* 121:1, pp. 148-177.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Scanlon, T. M. 2003. *The Difficulty of Tolerance: Essays in Political Philosophy*. Cambridge: Cambridge University Press.

Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford: Clarendon Press.

Scheffler, Samuel. 2010. *Equality and Tradition: Questions of Value in Moral and Political Theory*. Oxford: Oxford University Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, Massachusetts: Harvard University Press.

Schroeder, Mark. 2018. "Getting Perspective on Objective Reasons" in *Ethics* 128:2, pp. 289-319.

Schumm, George F. 1987. "Transitivity, Preference and Indifference" in *Philosophical Studies* 52:3, pp. 435-437.

Scruton, Roger. 2001. *The Meaning of Conservatism* (3rd Edition). Basingstoke, Hampshire: Palgrave.

Sen, Amartya. 2009. *The Idea of Justice*. London: Allen Lane.

Shafer-Landau, Russ. 1995. "Specifying Absolute Rights" in *Arizona Law Review* 37, pp. 209-224.

Shapiro, Ian, Rogers M. Smith, and Tarek E. Masoud (eds.). 2004. *Problems and Methods in the Study of Politics*. Cambridge: Cambridge University Press.

Sher, George. 1997. *Beyond Neutrality: Perfectionism and Politics*. Cambridge: Cambridge University Press.

Singer, Peter. 1999. "The Singer Solution to World Poverty" in *The New York Times Magazine* 05 September 1999, pp. 60-63.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Sleat, Matthew. 2013. *Liberal Realism: A realist theory of liberal politics*. Manchester: Manchester University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

Sobel, David, Peter Vallentyne, and Steven Wall (eds.). 2015. *Oxford Studies in Political Philosophy* (Volume 1). Oxford: Oxford University Press.

Sobel, Jordan Howard. 1997. "Cyclical Preferences and World Bayesianism" in *Philosophy of Science* 64:1, pp. 42-73.

Stark, Cynthia A. 2000. "Hypothetical Consent and Justification" in *The Journal of Philosophy* 97:6, pp. 313-334.

Steiner, Hillel. 2013. "Directed Duties and Inalienable Rights" in *Ethics* 123:2, pp. 230-244.

Stemplowska, Zofia. 2008. "What's Ideal About Ideal Theory?" in *Social Theory and Practice* 34:3, pp. 319-340.

Stilz, Anna. 2009. *Liberal Loyalty: Freedom, Obligation, and the State*. Princeton, New Jersey: Princeton University Press.

Temkin, Larry S. 1996. "A Continuum Argument for Intransitivity" in *Philosophy & Public Affairs* 25:3, pp. 175-210.

Thoma, Johanna. 2015. "Bargaining and the impartiality of the social contract" in *Philosophical Studies* 172:12, pp. 3335-3355.

Thomson, Judith Jarvis. 1977. "Some Ruminations on Rights" in *Arizona Law Review* 19, pp. 45-60.

Thomson, Judith Jarvis. 1986. *Rights, Restitution, and Risk: Essays in Moral Theory*. Cambridge, Massachusetts: Harvard University Press.

Thrasher, John. 2014. "Uniqueness and symmetry in bargaining theories of justice" in *Philosophical Studies* 167:3, pp. 683-699.

Thrasher, John. 2019. "Constructivism, representation, and stability: path-dependence in public reason theories of justice" in *Synthese* 196:1, pp. 429-450.

Ullmann-Margalit, Edna (ed.). 1988. *Science in Reflection* (The Israeli Colloquium: Studies in History, Philosophy, and Sociology of Science, Volume 3). Dordrecht: Kluwer Academic Publishers.

Vallentyne, Peter (ed.). 1991a. *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement. Cambridge: Cambridge University Press.

Vallentyne, Peter. 1991b. "Contractarianism and the assumption of mutual unconcern" in *Contractarianism and Rational Choice: Essays on David Gauthier's* Morals By Agreement, ed. Peter Vallentyne. Cambridge: Cambridge University Press.

Vallier, Kevin. 2013. "Can liberal perfectionism justify religious toleration? Wall on promoting and respecting" in *Philosophical Studies* 162:3, pp. 645-664.

Vallier, Kevin. 2014. *Liberal Politics and Public Faith: Beyond Separation*. London: Routledge.

Vallier, Kevin. 2016. "In Defence of Intelligible Reasons in Public Justification" in *The Philosophical Quarterly* 66:264, pp. 596-616.

Vallier, Kevin. 2018. "Public Justification" in *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), ed. Edward N. Zalta; https://plato.stanford.edu/entries/justification-public/ (Accessed 16 October 2018.)

Vallier, Kevin. 2019. *Must Politics Be War?: Restoring Our Trust in the Open*

Society. Oxford: Oxford University Press.

Vanderschraaf, Peter. 2019. *Strategic Justice: Convention and Problems of Balacing Divergent Interests*. Oxford: Oxford University Press.

Van Donselaar, Gijs. 2013. "Sticks or Carrots? The Emergence of Self-Ownership" in *Ethics* 123:4, pp. 700-716.

Vincent, Andrew. 1992. *Modern Political Ideologies*. Oxford: Blackwell.

Von Neumann, John and Oskar Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton, New Jersey: Princeton University Press.

Waldron, Jeremy. 1987. "Theoretical Foundations of Liberalism" in *The Philosophical Quarterly* 37:147, pp. 127-150.

Weale, Albert. 2013. *Democratic Justice and the Social Contract*. Oxford: Oxford University Press.

Wellman, Christopher H. 1996. "Liberalism, Samaritanism, and Political Legitimacy" in *Philosophy & Public Affairs* 25:3, pp. 211-237.

Widerquist, Karl and Grant S. McCall. 2017. *Prehistoric Myths in Modern Political Philosophy*. Edinburgh: Edinburgh University Press.

Williams, Bernard. 1981. *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press.

Williams, Bernard. 2005. *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton, New Jersey: Princeton University Press.

Zellentin, Alexa. 2012. *Liberal Neutrality: Treating Citizens as Free and Equal*. Berlin: De Gruyter.

Zeuthen, Frederik. 1930. *Problems of Monopoly and Economic Warfare*. London: George Routledge & Sons, Ltd.