DELIBERATION IN ANARCHY

by

Jonathan Michael Tomm

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

in
The Faculty of Graduate and Postdoctoral Studies

(Political Science)

The University of British Columbia

(Vancouver)

October 2014

© Jonathan Michael Tomm, 2014
Abstract

How do cooperative rules and agreements emerge in anarchical situations of political conflict between or within states? This dissertation responds to this question by developing an innovative theory of the role that deliberation plays in the emergence of rule-based cooperation under anarchy. “Deliberation” refers to a kind of political talk in which participants attempt to persuade each other on the basis of convincing reasons. Deliberation promises to be an important vehicle through which political opponents make and follow cooperative rules together. The problem, however, is that because deliberation has primarily been studied in the context of strong, liberal-democratic states, we do not yet have an adequate understanding of how it can be effective under more anarchical conditions. Without the security provided by strong institutions and relatively thick “lifeworlds” of shared culture and norms, deliberative responses to politics are more likely to be scuttled by mistrust and mutual suspicion.

The theory developed here addresses this challenge by highlighting and investigating the often-overlooked relational effects of deliberation. Deliberation is not just about the content of the reasons political opponents offer one another. It is also about the fact that they are offering reasons at all, and about what that implies for the kind of actions they can expect from each other. Most importantly, deliberation can be a way for interlocutors to demonstrate their accountability to each other: their willingness to uphold, and count on others to uphold, shared standards of behaviour. In this way deliberation can help to generate the bonds of trust and mutually acknowledged commitment that political opponents need in order to act on shared rules and understandings. In developing this theory, the dissertation offers new conceptual tools for
understanding how productive politics, based on negotiation, persuasion, and shared rules, can get a toehold even under difficult conditions.
Preface

This dissertation is original, unpublished work by the author, Jonathan Tomm.
Table of Contents

Abstract.............................................................................................................................................. ii

Preface................................................................................................................................................ iv

Table of Contents .................................................................................................................................... v

List of Tables .......................................................................................................................................... x

List of Figures......................................................................................................................................... xi

Acknowledgements .......................................................................................................................... xii

Dedication ............................................................................................................................................... xiii

Chapter 1  Introduction .......................................................................................................................... 1

1.1 Shared rules and cooperation in Northern Ireland ......................................................................... 1

1.2 Rebuilding the ship at sea .................................................................................................................. 4

1.3 A deliberative response .................................................................................................................... 8

1.4 The need for better theory .............................................................................................................. 13

1.5 Deliberation: part of a larger story ................................................................................................. 17

1.6 The work of theory-building .......................................................................................................... 18

1.7 A summary of the argument ........................................................................................................ 19

Chapter 2  Making Shared Rules in Anarchy: Rebuilding the Ship at Sea ........................................ 25

2.1 Shared rules, commitments, trust, and institutions ........................................................................ 27

2.2 Collective action problems and the need for rules ........................................................................ 32

2.2.1 Political conflict ......................................................................................................................... 33

2.2.2 The possibility of mutually beneficial cooperation ..................................................................... 34

2.2.3 Incentive problems .................................................................................................................... 36

2.2.4 Thinking of rules as tools: some caveats .................................................................................. 42
2.3 Difficult circumstances for rule-making

2.3.1 Weak governing institutions

2.3.2 Thin social norms

2.3.3 Rebuilding the ship at sea

2.4 Conclusion

Chapter 3  Deliberation, Trust, and Anarchy

3.1 What is deliberation?

3.1.1 Deliberation and emotion

3.1.2 Deliberation and bargaining

3.2 The problem of anarchy in deliberative theory

3.2.1 Deliberation and trust

3.2.2 Deliberation in anarchy

3.2.3 Foreshowing a response: rebuilding the ship at sea

3.3 The problem of coercion

3.3.1 Deliberative moments in conflict resolution

3.3.2 Coercion that supports deliberation

3.3.3 Coercion in a deliberative system

3.3.4 Deliberation, but is it democratic?

3.4 Conclusion

Chapter 4  Power and Persuasion: Rational Choice and Sociological Approaches to Understanding Cooperation

4.1 Rational choice approaches to rule-based cooperation

4.1.1 Bargaining

4.1.2 Cooperation and enforcement
4.1.3 Trust......................................................................................................................... 95

4.2 Power and persuasion: the limitations of rational choice approaches .................. 96

4.3 “Sociological” explanations: Culture, norms, and identity...................................... 102

4.4 Cooperation and the emergence of social order...................................................... 106

4.5 Conclusion .................................................................................................................. 116

Chapter 5 Acting for a Reason..........................................................................................118

5.1 A two-part conception of acting for a reason......................................................... 121

5.1.1 Motivation and reasons.......................................................................................... 123

5.1.2 The unity of practical reason................................................................................ 124

5.2 Rationality and utility maximization ........................................................................ 127

5.2.1 The assumption of harmonized ends................................................................. 127

5.2.2 Inner pluralism and rationality as integration..................................................... 135

5.3 Rationality and collective action problems............................................................... 138

5.4 Conclusion .................................................................................................................. 143

Chapter 6 Public Reasons and Collective Actions .........................................................145

6.1 The idea of a public reason........................................................................................ 147

6.2 Reasons and accountability....................................................................................... 151

6.2.1 Overlapping reasons are not sufficient for publicity........................................... 153

6.2.2 Overlapping reasons are not necessary for publicity......................................... 156

6.3 Public reasons and collective action........................................................................ 159

6.3.1 Comparison to game theory................................................................................ 161

6.3.2 Public reasons and strategic interaction............................................................. 163

6.3.3 Instrumental and normativized interactions....................................................... 166

6.4 The conditional validity of public reasons............................................................... 169
6.4.1 Understanding failures of collective action ............................................................ 171

6.5 Relationships and behavioural expectations ............................................................ 173

6.5.1 Experimental support ......................................................................................... 177

6.6 Conclusion ............................................................................................................ 179

Chapter 7 Public Reasons and Deliberation ................................................................. 180

7.1 Communication as deontic scorekeeping ............................................................... 183

7.1.1 Meaning is determined by structures of inference ........................................... 184

7.1.2 Inference as a social practice ............................................................................. 187

7.2 Deliberation and rational choice .......................................................................... 191

7.2.1 Cheap talk and costly signalling ....................................................................... 191

7.2.2 Inference versus information ........................................................................... 194

7.3 Deliberation and public reasons ............................................................................ 198

7.3.1 The problem of judgment aggregation ............................................................. 199

7.3.2 The problem of incentives ............................................................................... 202

7.3.3 A problematic answer: deliberative preferences .............................................. 207

7.3.4 A better answer: public reasons ....................................................................... 210

7.4 The contingency of public reasons in communication ......................................... 213

7.4.1 The essential normativity of speech and thought ............................................. 215

7.4.2 The contingency of normativized interaction .................................................. 218

7.5 Conclusion ............................................................................................................ 220

Chapter 8 A Theory of Deliberative Relationship-Building ........................................ 221

8.1 Commitment and perception of the other ............................................................. 223

8.2 A baseline of stable, non-deliberative interaction............................................... 229

8.3 Deliberative possibilities for change ..................................................................... 233
8.3.1 Deliberative possibilities for changes in perception ............................................. 234
8.3.2 Deliberative possibilities for changes in commitment ........................................... 237
8.3.3 The relative difficulty of changing commitment and perception ................................. 241
8.4 Deliberation and collective action: virtuous and vicious circles .................................. 244
  8.4.1 The circular dynamics of relationship-building .................................................... 244
  8.4.2 The virtuous circle ................................................................................................. 246
  8.4.3 The vicious circle ................................................................................................. 248
8.5 Reasons and relationships in collective will-formation .............................................. 250
8.6 Conclusion ............................................................................................................... 254

Chapter 9  Will and Coercion in Institutions for Collective Action ............................... 255
  9.1 The difference between will and coercion ............................................................... 258
  9.2 Moral core, coercive periphery .................................................................................. 261
    9.2.1 Incentivizing cooperation ..................................................................................... 263
    9.2.2 Backstopping cooperation ................................................................................... 268
  9.3 Coercion affecting will-formation: spillover and crowding out ................................. 269
    9.3.1 Spillover ............................................................................................................... 270
    9.3.2 Crowding out ...................................................................................................... 272
  9.4 Will-formation affecting coercion: institution-building ............................................ 273
  9.5 Conclusion ............................................................................................................... 275

Chapter 10 Conclusion: The Implications of Constituting a Public ............................... 277
  10.1 The constitution of a public: wider implications .................................................... 279
  10.2 Some final thoughts on democratic theory ............................................................ 283

Bibliography ............................................................................................................... 286
List of Tables

Table 7.1. The “doctrinal paradox”................................................................. 201

Table 8.1. Combinations of commitment and perception................................. 227
List of Figures

Figure 6.1. The Prisoners’ Dilemma ................................................................. 165
Figure 8.1. Components of stable cooperation .................................................. 222
Figure 8.2. Baseline interaction ......................................................................... 231
Figure 8.3. Virtuous circle .................................................................................. 247
Figure 8.4. Vicious circle .................................................................................... 248
Figure 9.1. The “moral core, coercive periphery” model .................................... 263
Figure 9.2. Deliberative and coercive modes of influence ................................... 264
Figure 9.3. Coercion affecting will-formation: spillover and crowding out ........... 270
Figure 9.4. Will-formation affecting coercion: institution-building .................... 274
Acknowledgements

This dissertation is the product of many years of learning, exploring, and thinking, both at the University of British Columbia, and in earlier study at the University of Victoria and Regent College, Vancouver. Through all those years I have been inspired, encouraged, and guided by many people. I am deeply grateful to them all.

Studying political science at UBC has been a richly rewarding experience. I am particularly grateful to my supervisor, Mark Warren. Mark has been the best supervisor one could ask for. He’s offered me endless advice and encouragement, helped me to frame (and re-frame, and re-frame again) the amorphous intuitions out of which this dissertation developed, read paper and chapter drafts with care and insight, engineered opportunities for me to make valuable contacts, and given me friendly and down-to-earth guidance through the many new experiences of a budding academic.

I wish to thank the other members of my advisory committee—Laura Janara and Dick Price—for their insightful comments and helpful suggestions. Even before I got to the dissertation, Laura and Dick inspired me as superb teachers of seminar classes, encouraged me toward an academic career, and offered me excellent advice.

One of the best things about grad school is the friendship and camaraderie of fellow students. I especially want to thank Agustín Goenaga, Chris Tenove, and Stewart Prest for all the helpful feedback they have given me through our semi-regular meetings over the last two years. Thanks to Aubin Calvert and Sean Gray for their insightful comments and ideas, especially in the early, formative years of this project. Thanks also to Michael MacKenzie for generous feedback on my writing, for many stimulating conversations, and for the pleasure of teaching together.

I owe a debt of gratitude to the Social Sciences and Humanities Research Council of Canada (SSHRC) for generously providing three years of funding during my PhD research.

Special thanks are due to my parents, Bill and Elaine, for all their love and unwavering support, to my brothers, Matthew and Kevin, for their abiding friendship and our many philosophical conversations, and to my uncle, William, for sparking in me a love for ideas. Finally, my most heartfelt thanks go to Sarah Nykolaishen, for everything.
to Sarah
Chapter 1  Introduction

In conflicts between or within states, much depends on the ability of the parties to arrive at agreed-upon rules, principles, or understandings. Such rules can take many shapes, from formal treaties to informal “rules of the game” and implicit understandings. But whatever shape they take, and whatever their particular content, their essential characteristics are that they limit mutually harmful actions, channel conflict into less destructive forms, and allow parties to achieve common benefits. Establishing shared rules marks the most basic and fundamental shift that can take place in the pattern of interaction between any group of political opponents. It is the central dynamic in the sort of virtuous circles or upward spirals that characterize productive political interactions and conflict resolution. Likewise, the erosion of shared rules is a central dynamic in the downward spirals by which conflicts deepen, the political space for compromise and negotiation shrinks, and cooperation falls apart.

1.1  Shared rules and cooperation in Northern Ireland

The conflict in Northern Ireland illustrates just how much is at stake in the existence of shared rules among political antagonists. The late 1960s and early 1970s saw many of the rules and implicit understandings that governed relations between unionists and nationalists fall apart. The result was a downward spiral of increasing violence. Early clashes took the form of nationalist demonstrations demanding civil rights for Catholics, and counter-demonstrations by unionists who saw the civil rights movement as a front for Irish nationalism and a threat to Protestant rule. Through 1968 and 1969, the clashes became increasingly violent, culminating in “the Battle of the Bogside” in August 1969: a riot in Bogside, Derry, that lasted three days and saw intense
fighting between the largely Catholic residents, the Royal Ulster Constabulary, and loyalist mobs. ¹ Eight people were killed and 1,800 families fled their homes. ² The riots, together with the introduction of internment without trial in 1971 and the Bloody Sunday massacre in January 1972, greatly increased Catholics’ sense of insecurity and grievance, and fuelled support for a revived Irish Republican Army (IRA). Unionists, too, felt increasingly threatened, and loyalist paramilitary violence rose, motivated in part by a will to demonstrate to the British and to Irish nationalists that “political appeasement of republicanism was not a violence-free option.”³ From 1969 to 1972, the number of deaths per year (of civilians, security forces, and paramilitaries on all sides) rose rapidly from 19 to 497.⁴

By the 1990s, however, the conflict had begun to take a different turn. There was growing acceptance among both nationalists and unionists of the need for a negotiated settlement. Within Sinn Féin and the IRA, a long process of conceptual change had begun to destabilize the original republican view of the conflict as a contest between an undivided Irish nation and British imperialism—a view that simply did not acknowledge the existence of unionists as a legitimate community with the right not to be forcibly incorporated into a united Ireland.⁵ This viewpoint had begun to shift, led particularly by a younger generation of leaders, toward acknowledgement of the “principle of consent”: the notion that any change in the constitutional status of Northern Ireland would have to win the consent of a majority in Northern Ireland.

---

¹ In the literature on Northern Ireland, “unionist” and “nationalist” generally refer to the more moderate parties, while “loyalist” and “republican” refer, respectively, to violent groups on the unionist and nationalist sides.

² McKittrick and McVea 2001, 56.

³ Mulholland 2002, 89.

⁴ McKittrick and McVea 2001, 325.

⁵ Farren and Mulvihill 2000, 104.
Ireland, and not simply a majority on the island as a whole.\textsuperscript{6} At the same time, unionists were doing their own recalculations. The demographic trend of a shrinking Protestant majority and, following the 1994 IRA ceasefire, the likely emergence of a broad nationalist alliance convinced many unionists that a negotiated settlement was a better bet than continued intransigence.\textsuperscript{7}

With this, unionists and nationalists had begun the process of establishing (and to some extent re-establishing) shared rules. As the peace process gained momentum, it became increasingly clear that the implicit “rules of the game” had changed. Majority opinion within both communities no longer condoned the political use of violence; instead, there was a broad cross-community commitment to ending the conflict. The existence of this normative shift was underlined when the 1998 Omagh bombing, perpetrated by a faction of the IRA, was roundly condemned by both unionist and republican leaders, and prompted renewed efforts at peace-making rather than the return to violence that many feared it would.\textsuperscript{8}

The emergence of new cross-community norms against political violence went along with efforts by leaders on all sides of the conflict (unionists and nationalists within Northern Ireland, as well as the British and Irish governments) to build a structure of shared principles and understandings on which a peace deal could ultimately be reached. Through a long series of dialogues, joint declarations, and agreements—including the Anglo-Irish Agreement of 1985, the Hume-Adams talks, secret back-channel talks between the British government and the IRA,\textsuperscript{9} the 1993 British-Irish Downing Street Declaration, the multiparty talks leading to the 1998 Good

\textsuperscript{7} Mulholland 2002, 133–35, 139–41.
\textsuperscript{8} Mulholland 2002, 149.
\textsuperscript{9} Dochartaigh 2011.
Friday Agreement, and the 2006 St. Andrews Agreement—the conflicting parties established, slowly and incrementally, a fabric of mutually acknowledged rules and commitments, together with the trust needed to abide by them.

The result has been in many ways the reverse of the downward spiral of the late 1960s and early 1970s. Whereas then norms of restraint and non-violence unravelled, through the 1990s those norms were reaffirmed. Whereas then mistrust and fear worked in a vicious circle, the 1990s saw the emergence of virtuous circles, in which small steps of trust and cooperation enabled further trust and further cooperation—including, ultimately, direct collaboration in government between the Democratic Unionist Party and Sinn Féin, groups who represented the opposed “extremes” throughout the conflict. While one should be careful not to draw too rosy a picture of community relations in Northern Ireland today, it is undeniable that interactions between the two communities have been fundamentally transformed. A conflict that at times approached a Hobbesian “war of all against all” is now shaped by a thick set of shared principles, rules, and institutions.

1.2 Rebuilding the ship at sea

This dissertation is about how those kinds of transformations come about. More specifically, it is about the emergence of shared rules and cooperation in relatively anarchical situations like the conflict in Northern Ireland. To say that Northern Ireland at the height of the Troubles was relatively anarchical is to point to two conditions: the weakness of state institutions and the breakdown of shared norms and cultural consensus. Institutional weakness was apparent in the

fact that, for most of the conflict, there was no state that commanded the allegiance of the whole population—naturally, since the identity and sovereignty of the Northern Irish state was itself the main object of the conflict. Moreover, the state struggled to provide some of the most basic goods that we take for granted in peaceful, liberal-democratic regimes, including protection from large-scale political violence. At the same time, the normative and cultural fabric of Northern Irish society was deeply divided along communal lines, and shared norms that existed prior to the outbreak of the conflict became significantly degraded through countless cycles of inter-communal violence and retribution.

These two features—weak governing institutions and thin fabrics of social norms—are present to varying degrees in a wide range of circumstances, not only in divided societies and civil conflicts like Northern Ireland, but also in weak states, transitional democracies, and international or global politics. Relatively anarchical circumstances like these are important because they have a big impact on how difficult or easy it is for further shared rules to emerge. The emergence of new shared rules and understandings—the kind of transformation we saw in Northern Ireland—is more difficult the fewer, and weaker, rules and shared institutions are to begin with. One of the things strong institutions and social norms do is facilitate the creation of new rules and cooperative schemes. By contrast, staring out at the “thin” end is harder. Here, political actors face a challenge that can be described as “rebuilding the ship at sea.” In relatively anarchical circumstances, political opponents who wish to establish shared rules and

---

11 Until 1972, Northern Ireland, while constitutionally part of the United Kingdom, had a devolved government housed at Stormont in Belfast. In 1972, direct rule from London was imposed and, apart from a failed attempt to return to Northern Irish government in 1974, remained in place until 1999, when the institutions of the Good Friday Agreement were put in place. Direct rule was once again imposed in 2002, with devolved government finally restored in 2007.

mutually beneficial institutions have to do so “at sea”—away from the protections and supports provided, under more benign conditions, by the strong institutions and normative cultures of well-governed, constitutional states.

Throughout the dissertation, I use the word “anarchy” as a shorthand for this complex of institutional weakness and thin social fabrics. I do not use “anarchy,” as it is often used, as a shorthand for “international relations.” As a matter of fact, the domain of international relations is, to a large extent, characterized by the twin conditions of institutional weakness and thin social norms. And these two factors do arguably capture the conceptual essence of the term “anarchy” as it is commonly used both in the discipline of International Relations and in political science generally. But my primary concern is with these two conditions themselves and with the systematic challenges they pose to cooperation and (as I will argue below) deliberation, regardless of whether they occur inside or outside a state.

Moreover, anarchy as I use the term here is a matter of degree. The institutions that support cooperation and shared rules can be stronger or weaker, so their strength cannot be a binary variable corresponding to the divide between the domestic and the international. Even within international relations, the strength of institutions for making and enforcing rules varies widely across different issue areas or regimes and among different sets of actors, as does the relative thickness of the social norms and shared cultures that also support cooperation. The same is true of states: there are many degrees of more and less anarchical situations of conflict within states, from strong, peaceful, constitutional states at one end to failed states and outright civil war at the other. But despite this great diversity, we can generalize about these situations on

the basis of the common characteristics they share. Wherever they are found, and in whatever degree, thin institutions and social norms pose the problem of “rebuilding the ship at sea”: of generating shared commitments and common rules without the safe haven of strong institutional enforcement and thick, pre-existing normative consensus.15

Why should we pay particular attention to such anarchical circumstances? There are both practical and theoretical reasons to do so. In practical terms, it is arguably where rules and shared commitments are the thinnest that they also have the capacity to do the most good. As I argued above, shared rules are an important way of securing common goods in the face of collective action problems, and the goods that are at stake in anarchical circumstances are often of immense importance. That is perhaps most obviously true where the stakes of overcoming a collective action problem can be the difference between ending and continuing violence. That is not to say that the goods at stake in more institutionalized contexts are not also important. Canadian citizens, for example, could benefit immensely if they could overcome collective action problems that stand in the way of a more comprehensive overhaul of the health care system. But in comparison to settling a violent ethnic conflict, establishing rules of war, or securing a stable democratic constitution in a weak, democratizing state, the stakes of rule-emergence in more anarchical situations are likely to be higher.

From a theoretical perspective, relatively anarchical circumstances give us the opportunity to sharpen our concepts of cooperation and institution-building by forcing us to look at tough cases that cannot be easily explained. Cooperation in the context of strong institutions is

15 Note that the problem of founding cooperative rules in anarchical situations is not only a concern of the present. It was also the central problem to which many early modern political theorists were responding. Both Machiavelli (1998; 1996) and Hobbes (1994) were concerned with how to found stable political orders in relatively anarchical circumstances, Machiavelli in the political disorder of Renaissance Italy, Hobbes in the aftermath of the English Civil War. See also Arendt 1965; Honig 1991. (My thanks to Laura Janara for this point.)
overdetermined.\textsuperscript{16} Where there are effective institutional sanctions and monitoring, it is easy—perhaps too easy—to answer the question of why people cooperate: because they know they will be punished if they do not. Similarly, where institutions are strongly legitimate—that is, where they have the support and loyalty of their members—it is also easy, and perhaps too easy, to explain why people cooperate: people buy in to the rules normatively, and so are intrinsically motivated to uphold them. Where neither of these boons is present (neither effective monitoring and sanctions nor normative consensus) rule-based cooperation is less likely—but perhaps more instructive when it does occur.

1.3 A deliberative response

My goal in this dissertation—and my response to this problem—is to develop a theory of the role that \textit{deliberation} plays in the emergence of rule-based cooperation in anarchy. “Deliberation” here refers to a particular kind of political talk: talk in which participants influence each other through what Jürgen Habermas calls “the forceless force of the better argument.”\textsuperscript{17} All politics involves influence or power. Deliberation is defined by a special form of influence: non-coercive influence based on the binding or compelling quality of good reasons.\textsuperscript{18}

Why focus on deliberation? There are two reasons. The first is the observation that, one way or another, deliberation seems to be crucial to the emergence of shared rules and cooperation in anarchical situations. What reason is there to think that? To begin with, the sheer ubiquity of argument in politics lends at least an initial plausibility to the claim that deliberation

\begin{footnotesize}
\textsuperscript{16} Heath 2001a, 156.
\textsuperscript{17} Habermas 1999, 450.
\textsuperscript{18} Warren 2002, 180–82.
\end{footnotesize}
is in some way central to the business of coordinating collective action and achieving common benefits. Nearly every example of cooperative political action appears to be surrounded by argument—and not only in strong, liberal-democratic states where “talk-based” politics is institutionalized and protected,\(^{19}\) but also in tough cases of divided societies, civil conflicts, and international politics.\(^{20}\) The practice of Western diplomacy, particularly as it has developed since the 19\(^{th}\)-century Concert of Europe, embodies the intuition that rules and cooperation can only be maintained if dialogue, negotiation, and compromise are privileged over force and deception.\(^{21}\) Efforts at conflict resolution—whether in Palestine, in Colombia, or Ukraine—inevitably involve calls for “dialogue.” And the intuition that talking is essential to establishing cooperation even finds support in laboratory experiments: over fifty years of experiments demonstrate that allowing people to talk to each other is the single most important condition enabling cooperation in Prisoners’ Dilemmas and related problems.\(^{22}\)

Of course, not all this talk is deliberation. When people talk, they can also engage in manipulation, deceit, moral brow-beating, purely strategic bargaining, and threats. Nevertheless, at least at first glance, much of the talk that surrounds successful cooperation and the emergence of new rules appears to take the form of argument and more or less genuine attempts to persuade. While it is often difficult for observers to declare, in particular cases, whether persuasion and the “force of the better argument” were decisive,\(^{23}\) there is at least a prima facie case to be made that they often are. If offering each other reasons is not a plausible way of coordinating mutually

\(^{19}\) Warren 2012a.


\(^{21}\) Nicolson 1988; Mitzen 2013; Schroeder 1994.

\(^{22}\) Balliet 2010; Sally 1995.

\(^{23}\) Deitelhoff and Müller 2005; Bächtiger et al. 2010.
beneficial action, most people who end up cooperating successfully seem, at any rate, to think it is.\(^\text{24}\)

Moreover, there are reasons to think that deliberation is particularly important in relatively anarchical circumstances. To create new shared commitments and rules, particularly where participants start out with little in common, requires something that Hannah Arendt called “natality”: the ability to bring something new into the world.\(^\text{25}\) Deliberation has just this sort of generativity: the capacity to create new intersubjective understandings. The give and take of reasons seems uniquely suited to establishing new shared understandings of common problems and possible solutions.

To get a sense of the promise of deliberation in this regard, it is useful to compare it to the types of communication that can be modelled strategically, such as “cheap talk” and “costly signalling.” In rational choice models, “cheap talk” refers to communication by which participants signal which strategy they will play in a game with more than one equilibrium strategy.\(^\text{26}\) “Costly signalling” refers to communication in which players signal their “type” (which is usually a feature of their preferences) to each other by taking an action that incurs a cost only a player of that type would incur.\(^\text{27}\) For example, one state might signal to another state that it has “status quo” rather than “revisionist” preferences by withdrawing from territory that would give it a military advantage in the event of a conflict.\(^\text{28}\) What these types of signalling

\(^{24}\) Risse and Kleine 2010.

\(^{25}\) Arendt 1998.

\(^{26}\) Such talk is referred to as “cheap” because it does not cost the players anything to send the signal. As a result, cheap talk loses its meaning in situations where players have asymmetrical payoffs in different equilibria. See Chapter 7.

\(^{27}\) Morrow 1999.

\(^{28}\) Kydd 2000. Kydd uses the types “mean” and “nice.”
have in common is that what can be *communicated* in such models is limited to a set of possible beliefs, “common priors,” that participants must already share (and know that they share) before they start talking.\(^{29}\) Thus, one state might be able to signal to another whether it is “status quo” or “revisionist,” but only because both states already know that those are the two options.\(^{30}\)

Deliberation, though, is not limited to shifting the preponderance of evidence among various pre-existing claims that are fully known to everyone at the outset.\(^{31}\) Anyone who has engaged in a genuinely open-ended argument with someone who sees the world differently enough knows that arguments can sometimes take unexpected turns, and may even arrive at new possibilities and new *meaning* that neither party had thought of before.\(^{32}\) That, indeed, is one reason that deliberative democrats insist on the necessity of actual conversation between citizens, and reject the “monological mock dialogue”\(^{33}\) of the philosopher or social scientist attempting to re-create the result of an ideal and hypothetical dialogue, however knowledgeable and just that person might be. Thus, in situations where the emergence of rule-based cooperation depends upon participants coming to see themselves, each other, or their common problems in new ways—ways they had not contemplated before—deliberation would appear to have a crucial role to play.

A final reason to look to deliberation is that it promises to fill a gap in our current understanding of how cooperation emerges in anarchical circumstances. As I will argue in

\(^{29}\) Kreps 1990, chap. 13.

\(^{30}\) For the details of this argument, see Chapter 7.

\(^{31}\) Warren 1999a.

\(^{32}\) As Habermas argues, reasons often turn out to have meaning that “overshoots” the intentions of individual speakers (1996, 11, 15).

\(^{33}\) Habermas 1987, 2:95.
Chapter 4, neither rational choice nor “sociological” approaches to explaining cooperation adequately address the challenges posed by anarchy. (By “sociological” I mean approaches that focus on the role of what Habermas calls the “lifeworld”—habit, shared meaning, and social identity—in fostering cooperation and trust.) On the one hand, rational choice approaches are very good at explaining the role and effects of power and strategic reasoning in cooperation. But they are much more limited when it comes to explaining how normative commitments and intrinsic motivations contribute to cooperation. (Indeed, one important contribution this dissertation makes is to offer a rich account of precisely why rational choice models are limited in this regard.) On the other hand, the difficulty facing sociological explanations of cooperation is that anarchical circumstances, as I have defined them here, are marked by a relative paucity of precisely the social resources—the shared norms, friendly identities, and cultural consensus—such explanations rely on. In relatively anarchical circumstances, the challenge is to explain where those social resources for cooperation come from. Deliberation offers a promising way to fill this gap. If deliberation does have the capacity to generate new intersubjective understandings and shared norms, it may help to explain how political opponents build the kind of normative commitments that tend to be assumed in sociological explanations of cooperation and overlooked in rational choice ones.

34 For the classic distinction between rational choice and “sociological” theories, see Barry 1970; Elster 1989a. “Sociological” is a somewhat problematic label, as it denotes a field of study rather than any particular theory or family of theories, but it still makes sense as a category for reasons I will explain in Chapter 4.

35 Habermas 1987.

36 See Chapters 6 and 7.
1.4 The need for better theory

But if it is true that deliberation is, or at least might be, crucial to the emergence of rule-based cooperation, it is also true that we do not yet have an adequate theory of how deliberation could play such a role. That need is the second fact that motivates this study. The idea of deliberation has seen tremendous development in the last twenty-five years, especially as a contribution to democratic theory. But the very fact that the theory of deliberation has developed in such close association with democracy, and especially democracy within the liberal-democratic state, has also left it with some blind spots when it comes to theorizing more anarchical situations.

Deliberative democrats have recognized that deliberation is a somewhat difficult, limited, and precarious way of coordinating action, particularly in the face of political conflict. In order to deliberate, people have to feel reasonably secure, and they have to feel confident that in communicating openly and abiding by their verbal commitments, they will not simply be suckered by less scrupulous opponents. Added to this is the fact that, even with the best of intentions, people’s resources for reaching mutual understanding only go so far. Where participants confront each other from deeply different perspectives and cultural assumptions, they may simply fail to find a way to come to an understanding.

Liberal democracies succeed in sustaining (relatively) deliberative politics by drawing on two important resources: law (backed up when necessary by coercion) and thick “lifeworlds” of

37 Dryzek (2010, chap. 1) places the beginning of the “deliberative turn” in democratic theory around 1990.

38 In the last twenty years or so (counting from the beginning of the “ZIB Debate” [Deitelhoff and Müller 2005, n. 1]), there has also been a significant development of the idea of deliberation, or “arguing,” in the IR literature. But the IR and democratic streams of deliberative theory have developed largely in isolation from each other.

background normative and cultural consensus. These democratic institutions and cultures do much to make deliberation safe, which is crucial if people are to rely on the force of arguments despite the mutual suspicion and antagonism that come naturally with political conflict. Indeed, strong institutions often work so well that we fail to notice just how much work they are doing.

But all that changes when deliberation takes place in more anarchical circumstances. And partly because deliberative democrats working in the context of strong, constitutional states have been able to take these securities for granted, we do not yet have as good a grip as we might on how deliberation could be effective in the absence of those state-based securities. Most importantly, we need to better understand how deliberation itself might stabilize cooperation. The key insight of deliberative theory is that reasons and reason-giving have their own motivational force, distinct from the force of the institutional sanctions and unquestioned social norms that sometimes stand behind them. Good reasons should be able to stabilize cooperation by creating relatively secure and reliable cooperative motivations among interlocutors who are convinced by them. But the connections between reasons and actual motivations can be weak. While deliberative democrats have long understood deliberation as a way of coming, in effect, to good answers about questions of public policy, we have not yet done a good enough job of understanding how deliberation might encourage people to act on those “good answers.” We need to better understand how deliberation contributes to those crucial moments when shared commitments coalesce—when political actors say, in effect, “okay, I’ll jump if you jump,” and then actually jump. What exactly is the nature of those collective commitments, and how do they achieve a measure of solidity, such that people can act on them despite the risks of doing so in relatively anarchical circumstances?

In the chapters that follow, I will present a theory of deliberation that helps to answer these questions. In particular, I will argue that the key to understanding how deliberation contributes to cooperation is recognizing its relational significance. Deliberative theorists are right that the give and take of reasons can give rise to shared commitments and shared actions—though only among people who relate to each other in a certain way. Specifically, the give and take of reasons, and the achievement of shared understandings, goals, and rules relevant to action, is something that only happens among people who relate to each other on a basis of mutual accountability. It is only where people take themselves to be responsible to each other for the rightness, truth, or “validity” of their claims that shared reasons have a genuine motivational grip on people, and are thus a reliable way of coordinating action. Deliberation is capable of generating new shared commitments because it is not just about the content of what is said; it is also (implicitly) about the people saying it, and specifically about their mutually acknowledged accountability to each other. The work that deliberation does is not only cognitive, in the sense of coming to the right answers, but also relational, in that it establishes the social bonds, solidarity, and intersubjective commitments needed to act on shared understandings. It is in highlighting and theorizing these often-overlooked relational effects of deliberation that my dissertation makes its main contribution.

If we can develop better theories of how deliberation works in the most difficult circumstances, the payoffs should be enormous. For one thing, we would have new conceptual tools to better understand how outcomes like the one in Northern Ireland happen: outcomes in which political interactions begin to shift away from self-defeating distrust and strategic defection toward more productive politics based on negotiation and persuasion. Such a theory should help us identify how this sort of productive, deliberative interaction can get a toehold in
such situations, and how vicious circles of tit-for-tat defection might be turned into virtuous circles, in which political opponents build an increasing capacity to act together on the basis of shared rules.

If deliberation can contribute to such virtuous circles in otherwise difficult conflicts, it clearly has normative value simply in terms of human welfare. But the normative value of deliberation also goes beyond its usefulness as a means of resolving conflict and cooperating for joint gains. One reason deliberation has attracted the broad attention it has among political theorists in recent years is that it embodies normative values that are deeply and widely held in modern (perhaps Western) societies. The ideal of deliberation contains within it implicit commitments to relationships of mutual recognition, equality, autonomy, and non-domination. As I will argue at more length later, the practice of giving and offering reasons to each other is freighted with mutual recognition—at the very least, recognition that one’s interlocutors are themselves capable of recognizing and abiding by good reasons. Likewise, there is an implicit equality bound up with reason-giving. Good reasons (if they are good reasons) bind everyone alike and, at least ideally, a good reason is compelling regardless of the status, wealth, or power of the person offering it. For these reasons, to the extent that collective decisions come to be mediated through speech and persuasion instead of pure violence, wealth, or status, they should exhibit a bias toward actions that respect the equality and autonomy of all participants, and away from actions based on unlegitimated and arbitrary power. That is not to say that all actual

---

41 Warren 2007, 274–75.
42 See, especially, Chapter 8.
43 This is part of Habermas’s (1984; 1987) theory of “the rationalization of the lifeworld.” See also Heath 2014a.
deliberation always embodies these ideals; it clearly does not. The point, rather, is that the ideals that are implicit in deliberation are consonant with a wide range of normative commitments (including liberalism and republicanism) in modern political theory.

1.5 Deliberation: part of a larger story

The promise of deliberative theory for understanding cooperation in anarchy is great. But that promise is most likely to be fulfilled if we are also aware of deliberation’s limitations. Those limitations are implicit in the very problem I am setting out to address: how cooperation emerges in the difficult circumstances created by institutional weakness and thin social norms. These are difficult circumstances precisely because institutions and background norms are themselves enormously important means of stabilizing cooperation.

Thus, while theorists of deliberation are committed to the notion that good reasons have a distinctive force—one that cannot be reduced to the coercive incentives and unquestioned social norms that sometimes stand behind them—they are not committed to the claim that the force of reasons is, or even should be, the only force in politics. Rather, deliberative theory captures a particular slice of concerted social action, namely, action that is voluntary and reason-based. That, in effect, is the “value added” of deliberative theory. By thinking about reasons—what it means to act for a reason, what happens when reasons are addressed by one person to another, and how such reasons then affect actions in situations of strategic interdependence—we gain important insight into how people might choose to adopt new rule-based schemes of cooperation.

44 See the discussion at the end of Chapter 3 on the prospects for democratic deliberation in relatively anarchical situations.

45 Hannah Arendt’s (1998; 2005) work suggests that speech and persuasion were central ideals in ancient Greek and Roman politics as well.
But voluntary, reasoned action is not the whole story. First, stable political systems are always a mix of the voluntary and the involuntary, of consent and coercion. And, second, even action that is largely voluntary is itself more or less rational—which is to say, shaped to a greater or lesser degree by conscious, reflective choice, as opposed to habit and unconscious (and thus unexamined) cultural assumptions. Thus, the deliberative account of cooperation that I am developing here should not be taken to deny the importance of other motivational forces and sources of social action. To the contrary, I argue that we need to better theorize the ways deliberative and non-deliberative forms of influence work together in establishing and maintaining cooperative arrangements over the long run.

1.6 The work of theory-building

Before summarizing the argument to come, I should clarify the contribution this dissertation will make. I see my dissertation as doing the work of theory-building. My goal is to develop a conceptual framework for thinking about how deliberation might contribute to the emergence of cooperation in relatively anarchical circumstances. The work here lies in conceiving of deliberative will-formation more clearly: of understanding what it would mean for the “force of the better argument” to be a means of coordinating action under relatively anarchical the conditions.

Such work needs to be empirically attentive, in that it needs to respond to what we know about how people make decisions, about how they respond to difficult circumstances and insecurity, about the cognitive and motivational limitations of rational or reflective decision-making, and about the way people assess each other’s motivations and intentions, among other
things. And like any theory, it should be insightful, in that it should illuminate or make sense of practices or phenomena that, without the theory, would remain opaque or mysterious.

But while the work of theory-building must be empirically attentive, it is a different task from the empirical testing of a theory. The dissertation is not structured to test the conception of deliberative will-formation developed here. In order to ask the basic question of empirical research, “how do I know this is right?,” we must first know what it would mean for deliberative theory to be “right” in these particular circumstances, and on that question, as I argued above, there is still much work to be done. My goal here is to develop the sort of broad conceptual framework that may then serve to generate more specific hypotheses of the kind that lend themselves to detailed case studies and other modes of empirical evaluation.

1.7 A summary of the argument

The argument of the dissertation will proceed as follows. Chapters 2 to 4 present the problem of the dissertation in more detail and situate it within the literatures on deliberation and cooperation. In many ways, these chapters are an extension of this introduction, but this extended introductory material is important because in this project setting up the problem clearly is half the battle. This dissertation deals with a very broad and basic problem of political life. Moreover, the domain of the inquiry—“relatively anarchical circumstances”—is similarly broad, encompassing a wide range of situations including weak states, transitional democracies, divided societies and civil wars, international and global politics, and even, conceivably, the wilder, less institutionalized, and more socially divisive reaches of politics within otherwise strong, liberal-democratic states. In order to talk sensibly about such a broad domain, it is necessary to be very specific on a theoretical level about what problems exactly are being addressed.
In Chapter 2, I explain the research problem in detail: how does rule-based cooperation emerge in relatively anarchical circumstances? There are two parts to this chapter. The first explains what I mean by “rule-based cooperation.” Cooperation is a nebulous term, one that is used in quite different ways. I sharpen the theoretical focus using the framework of collective action. I argue that political opponents build shared rules and cooperative schemes in response to collective action problems. That is, they use “rules as tools” to help them realize common interests in situations where they also face incentives to defect from cooperative schemes. The second task of the chapter is to explain the problematic context in which this collective rule-making takes place, namely, “anarchy.” I specify the conditions of institutional weakness and normative or cultural “thinness” that make contexts of interaction relatively anarchical. I then return to the metaphor of re-building the ship at sea to explain the difference these difficult circumstances make for cooperation.

In Chapter 3, I turn to deliberative theory in order to survey the resources it contains for addressing the problem of cooperation in anarchy. I argue that the concept of collective will-formation is the key to understanding how deliberation contributes to cooperation. Deliberative theorists, however, need to clarify how such collective will-formation is possible in light of the challenges posed by these difficult circumstances. I argue that the key to doing so is to recognize that collective will-formation has, in effect, two dimensions: it is about both the content of the reasons participants address to each other, as well as the relationships of mutual accountability that must obtain between those participants in order for the practice of giving and asking for reasons to make sense. As deliberative theory moves into increasingly anarchical circumstances,

\[\text{Ostrom 2005.}\]
its relationship-building potentials become all the more important, but we do not yet understand these relational implications of deliberation as well as we should.

In Chapter 4, I argue that deliberation has the potential to fill an explanatory gap in our current understanding of how cooperative rules emerge in anarchy. I divide the present literature into two main bodies: rational choice and “sociological.” Regarding rational choice approaches, I argue that they are very good at explaining the role and effects of power and strategic reasoning in cooperation, but that they are not adequate to explain the contribution of persuasion and normative commitment. Sociological approaches, by contrast, are at first glance well-suited to explaining norm- and commitment-based cooperation. But they are better at explaining trust and cooperation within groups than between groups, and better at explaining the long-term stability of cooperation than its initial emergence. The challenge for these theories is to say where social resources for cooperation—the friendly identities and cooperative norms—come from, particularly in situations of violent conflict where social bonds between groups are weak. Thus, neither approach can answer the most important question: how cooperation emerges in anarchy. Deliberative theory promises to fill this gap.

With that background in place, I then begin the task of developing a deliberative theory of rule-based cooperation. I start, in Chapters 5 to 7, by examining and untangling the key concepts on which the theory depends: practical rationality (Chapter 5), public reasons (Chapter 6), and deliberation (Chapter 7).

Chapter 5 examines the notion of practical rationality, or what it means to act for a reason. If deliberation is a form of influence through the give and take of reasons, and if we are interested in the effects of deliberation on actions—and ultimately on the emergence of rule-based cooperation among political opponents—we must first get clear on how reasons and
actions relate to each other. Specifically, it is essential to understand reasons as being inherently both *normatively prescriptive* and *motivationally effective*. A reason just is a motivation that an actor endorses as right or good. Thus, if an actor recognizes herself as having a reason to act, there is no need to posit an additional motivational mechanism (such as self-interest) to explain why she acts on that reason.

Chapter 6 builds on that conception of practical rationality to develop a better understanding of collective action. I argue that collective action depends upon the emergence of a certain kind of reason, namely, *public reasons*. Public reasons are reasons that are shared, but shared in a particular way. Reasons are public when they are addressed to each other as justifications. The validity of public reasons depends, however, upon the existence of a public, which is a group of people who consider themselves accountable to each other and whose reasons for acting are responsive to shared standards of justification. In collective action problems, decisive reasons in favour of cooperation tend to be public reasons. That means that whether people cooperate turns on whether they see themselves as forming a public—that is, whether they and the others with whom they are interacting see themselves as being accountable to each other. In this way, the challenge of collective action is ultimately the challenge of constituting a public.

Chapter 7 connects the notion of public reasons to deliberation. Reasons given in deliberation are necessarily public reasons—reasons, that is, that participants address to each other *as* reasons, as claims that affect what each is responsible for. Mutual accountability is not an ethical or moral requirement layered on top of deliberation. To the contrary, it is the vehicle of communication itself: people mean the same thing by a given set of words if they share judgments about what the speaker commits herself to by uttering those words. The reasoning that
is distinctive of deliberation as a form of communication—reasoning in which inferences, and not only information, are in play—is thus inherently a practice of holding each other accountable in normative terms, or what Robert Brandom calls “deontic scorekeeping.” Establishing this inherent connection between deliberation and mutual accountability is an important step toward explaining deliberative influence in anarchy. This mutual accountability is what gives deliberation the capacity to build the normative commitment and trust needed for cooperation.

In Chapter 8, the dissertation shifts gears. In this chapter, I begin using the conceptual resources developed in Chapters 5 to 7 to outline an explanatory account of how deliberation contributes to cooperation. The theory focuses on how political opponents come to see themselves as standing in relationships of mutual accountability to each other. I begin with individual-level attitudes of commitment to public reasons and perceptions of others’ commitment. From there I conceptualize relational steady states: relatively stable combinations of perception and commitment that result in either cooperation or defection. And from there, I survey the possibilities for relational change through deliberative and non-deliberative interactions, including both virtuous and vicious circles. The result is a theory of how opponents build shared normative commitments and mutual trust through acting and talking together.

Chapter 9 puts this conception of voluntary, reason-based cooperation into the broader picture of stable cooperative regimes. I argue that rule-based cooperation has a “moral core, coercive periphery” structure, in which largely voluntary cooperation is backed up, where necessary, by coercive institutions. Such institutions are crucial to the success of deliberative politics because they offer a needed measure of assurance that those who rely on public reason

47 Brandom 2000, 81.

and shared rules will not be suckered by others. One of the most important contributions deliberation can make is thus to foster the collective will needed to build and maintain such institutions—and in that sense, help to make anarchical circumstances less anarchical.

Chapter 10 summarizes the main findings of the dissertation and offers some reflections on their wider significance. While the dissertation takes as its starting point the problem of how cooperative rules emerge in relatively anarchical situations, the resulting theory is highly general in its applicability. In particular, the connections between *publicity* and *collective action* developed here have the potential to contribute to the study of collective action in a wide range of circumstances. I conclude, finally, with some reflections on what the theory might imply for the study of democracy.
Chapter 2  Making Shared Rules in Anarchy: Rebuilding the Ship at Sea

Northern Ireland is widely seen as a successful example of conflict resolution.\(^1\) Looking at the course of the conflict, there is clearly something “right” about the turn it took in the 1990s. There has been a significant transformation in the way politics works in Northern Ireland from the 1970s to today. Then, paramilitary violence was rife, the nationalist and Catholic population was largely alienated from the state (and especially its security services), and government within the province had effectively broken down.\(^2\) Now, by contrast, paramilitary violence has substantially ended, nationalists and Catholics are included and represented in state institutions (including a restructured and integrated police service), the “extreme” unionist and nationalist parties (the Democratic Unionist Party and Sinn Féin) collaborate in government, and there is a widespread popular sense that the violence of the past has been laid to rest.\(^3\) That is not to say, of course, that things could not be better: for one thing, the social division between unionists and nationalists in Northern Ireland remains deep.\(^4\) But from a big-picture perspective, and in comparison to other cases of violent civil conflict, Northern Ireland stands as a hopeful example.

\(^{1}\) There have been numerous attempts to apply “lessons” from the Northern Irish experience to other difficult conflicts. For example, in September 2007, a group of Northern Irish politicians were brought to Finland to talk with Iraqi politicians (Trimble 2007, 4). For some concerns about whether the right “lessons” are being drawn, see Trimble 2007; Mac Ginty 2011; O’Kane 2010.

\(^{2}\) After the collapse of the Sunningdale power-sharing executive in 1974, Northern Ireland was governed directly from London.

\(^{3}\) Confidence that the new power-sharing institutions will hold has grown slowly since their establishment in 1998. Initially, a significant portion of unionists rejected the Good Friday Agreement, but now, as Paul Bew (2007, 1) points out, “not one of the 108 members of the Stormont Assembly elected in March 2007 has a fundamental objection to the GFA template—only slightly revised by the St Andrews negotiations of October 2006.”

But how exactly should we characterize this transformation? How can we describe this sense of “things going right” in a way that could be generalized to other situations? Perhaps the most straightforward way to describe what happened is “cooperation.” Unionists and nationalists began to cooperate with each other. That is true, but it is also vague. One of the difficulties in generalizing about cooperation is that it can mean many things. “Cooperation” is part of a group of loosely-related terms, including “coordination,” “collective action,” and “following the rules.” And lying behind these terms is a similar profusion of theoretical frameworks and assumptions. Say “cooperation” to a game theorist and he or she might think about Nash equilibria and Pareto improvements. Say “cooperation” to an IR constructivist, and he or she might think about constructed “friend” identities and security communities.5

My goal in this chapter is to clarify our somewhat vague notions about what is “going right” in success stories like Northern Ireland, and thus to better define what it is that I hope to explain with the deliberative theory I will develop in the rest of the dissertation. I suggest that rather than theorizing “cooperation” in a broad sense, we should focus more specifically on the emergence of _shared rules_. The emergence of shared rules is what happens when political actors shift away from essentially “no holds barred” struggle toward conflict shaped by an increasingly thick fabric of shared rules, commitments, and mutually acknowledged obligations. Building and maintaining a structure of shared rules together is the _activity_ that makes that happen.

If defining that activity is the first task of this chapter, the second task is to specify in more detail the _difficult circumstances_ of institutional weakness and social division this dissertation will focus on. Building a structure of shared rules is one thing. Doing so under relatively anarchical conditions—in weak states, civil conflicts, divided societies, and

5 E.g. Wendt 1999; Adler and Barnett 1998.
international politics—is another. Describing those difficult conditions and specifying the difference they make to how, and whether, political actors can succeed in establishing new, rule-based systems of cooperation is crucial to clearly conceiving the problem this dissertation addresses.

I begin, in section 2.1, by setting out what I mean by “shared rules” or “rule-governed interactions.” I specify the relationships between rules, commitments, trust, and institutions. In section 2.2, I sharpen this conception by thinking of it from a functional perspective. Political opponents build shared rules in response to collective action problems; that is, they use rules as “tools” to help them realize shared goals in situations where individuals face incentives to defect from cooperative schemes.⁶ This perspective helps me distinguish shared rule- and commitment-making from other, less precise, conceptions of cooperation or coordination. Finally, in section 2.3, I turn from defining the activity of building shared rules to considering the relatively anarchical circumstances I am interested in. I describe what makes circumstances like civil wars, international politics, and weak states (relatively) anarchical and, most importantly, the difference these difficult conditions make for the activity of building shared rules.

2.1 Shared rules, commitments, trust, and institutions

This dissertation is about how shared rules emerge under relatively anarchical conditions. But what exactly are shared rules? My first task is to define shared rules, and to relate them to the closely associated concepts of commitments, trust, and institutions.

A shared rule is simply an “ought” in the broadest sense: any norm, principle, deal, or understanding that the actors involved recognize as binding on themselves and others. Rules constitute social order. They mark the difference between interactions that are, in essence, a free-for-all and interactions that are conditioned, to at least some degree, by shared standards of judgment, mutual obligations, and self-restraint.

Rules can take many forms. They can be implicit or explicit. Political actors who interact over a long time often develop a set of unspoken “rules of the game,” implicit understandings about what can and cannot be done in the pursuit of individual goals. Of course, rules can also emerge explicitly, through negotiation and argument. Rules can be more or less formal, where formal rules are codified in detailed treaties or other legal documents. Rules may or may not be embedded in organizational structures. They can be regulative or constitutive in content: that is, they can prescribe or prohibit actions (regulative) or define the meaning of shared terms, identities, or activities (constitutive). They can also be deontological or consequentialist in content, where deontological rules prescribe or prohibit particular actions and consequentialist rules prescribe or prohibit goals (Ostrom 2005). When they are used, all of these forms of rules represent a thickening of social order and a step away from unrestrained and unlegitimated coercion.

A commitment is similar to a rule. Negotiating a peace deal often involves making commitments to do things that are only done once, such as, to return to the Northern Irish example, the commitment Sinn Féin made, and eventually kept, to disarm the IRA.

---

7 Bull 1977.
9 Searle 1969, 33–42.
“Commitment” is a convenient way to refer to such one-off obligations, as distinguished from rules, which are usually understood to govern interactions over the long run. Either way, both rules and commitments have the same character of being “oughts,” obligations that, when jointly made or affirmed, the conflicting parties recognize as binding on themselves and each other.

Along with shared rules and commitments goes trust. For political opponents, the challenge of rule-making is partly about agreeing on what the rules or deals should be. But it is also about trusting each to follow through on whatever is agreed. Making and abiding by rules and commitments is risky. By following rules, political opponents make themselves vulnerable to each other’s opportunistic defection and this, of course, is one of the things that makes it so hard for conflicting parties to negotiate agreements.\(^\text{10}\) In anarchical situations, moreover, cooperation is all the more risky since institutions that might otherwise enforce rules are weak or altogether absent. Such risks can be reduced by arranging incentives to minimize the temptation to defect, and by building trust and cooperation together in small steps.\(^\text{11}\) But the risks of cooperation can never be wholly eliminated, and thus building and following rules together nearly always requires building trust as well.\(^\text{12}\)

Shared rules often compose or contribute to institutions. Despite its importance in the social sciences, there is little agreement about how to define the term “institution.”\(^\text{13}\) In the broadest sense, all socially constructed realities and shared meanings might be referred to as

\(^{10}\) On such “commitment problems” see, e.g., Fearon 1995; Walter 2009.

\(^{11}\) Kelman 2005; Kydd 2005.

\(^{12}\) Luhmann 1979; Baier 1986; Barbalet 2009.

\(^{13}\) Scharpf 1997, 38; Searle 2005; A. A. Stein 2008.
institutions. In this sense, money, private property, and friendship are institutions. Some scholars understand institutions primarily as systems of regulative rules, as “[the] humanly devised constraints that shape human interaction.” In other cases, we use the term “institution” to refer to organizations or corporate actors who administer, adjudicate, or enforce a system of rules or act in accordance with a set of goals or routines. In this sense, the Supreme Court is an institution, as is, for example, the International Monetary Fund.

For my purposes here, it is not necessary to settle the question of how exactly to define “institution.” What is important is that whatever institutions are, shared rules are essential to them. That goes especially for constitutive rules: rules that take the form “X counts as Y.”

Thus, the fact that the piece of paper in my hand counts as a $20 bill is an institutional fact, because it depends for its truth on our collective agreement that holding it confers a particular status on me, namely, the status of being entitled to exchange it for other goods of the same value. Likewise, the institution of state sovereignty is constituted by a rule (or a system of rules) that define a particular status, namely, the status of being immune from political interference and territorial incursion by other states. The key point is that these institutional rules, like all rules, have the basic character of “oughts.” In other words, institutional structures are “deontic

---

14 Searle 2005, 10, 18.
16 Many scholars would use the term “institution” for a broad system of rules, practices, and social beliefs, and “organization” to denote a corporate entity with the capacity to act (including such things as offices, personnel, equipment, and budgets). Institutions may or may not have such organizational expressions. Scharpf 1997; Adler and Barnett 1998.
17 As Bo Rothstein (1996, 145) suggests, “there seems to be a general agreement that, at their core, political institutions are ‘the rules of the game’.”
18 Searle 2005.
powers”: they are essentially composed of obligations, entitlements, permissions, rights, and other normative status people can assign to each other.\textsuperscript{20}

As a result, much of the rule-building that I am theorizing in this dissertation could also be described as institution-building. While institutions are often more than simply a set of rules—particularly when we use the term “institution” to refer to organizational entities with a capacity to act— institutions have rules at their core. As interactions between political opponents become more thickly institutionalized, they become increasingly constrained and constituted by shared standards of behaviour and shared “oughts”—that is, by rules in the broad way I use that term here.

Finally, before moving on I should clarify that to speak of “shared rules” is not to prejudge the issue of how actors are motivated with respect to rules. Rules are necessarily “oughts,” and in that sense are a normative phenomenon, but why a person complies with, or fails to comply with, these oughts is another question. Thus, to speak of rules is not necessarily to say that actors follow a “logic of appropriateness” rather than a “logic of consequences.”\textsuperscript{21} People can have all sorts of reasons for following rules, including both normative commitment and instrumental calculations about material costs and benefits. How people are motivated to uphold shared rules and mutual commitments is an important question, and one that I will explore in considerable detail in the chapters that follow, but no particular answer to that question has to be assumed in advance in order to grasp the problem my dissertation addresses.

\textsuperscript{20} Searle 2005, 10.

\textsuperscript{21} Cf. March and Olsen 1998.
2.2 Collective action problems and the need for rules

The concept of shared rules can be sharpened and clarified by thinking through the functions they serve—that is, what rules accomplish for the people who make and follow them together. I will argue that shared rules can be understood as a response to collective action problems. Such a functionalist account of rules does not tell the whole story, and I will have more to say about the limitations of this approach later. But it is an important start, and it does much to sharpen the concept of shared rules and place it in relation to less precise notions of cooperation.

Collective action problems (as I will define them here) arise when three conditions are in place: (1) political conflict, (2) common interests, and (3) incentive problems. On the functionalist account I’ll present here, these are the conditions under which political opponents will have a need for rules: the conditions under which rules will be useful to them. Thus, these conditions can also be understood as defining scope conditions for this study since, where any of these conditions is missing, opponents are likely to have neither the interest nor the opportunities for creating rules in the first place. But as I will also argue below, there are reasons to suspect that all three conditions obtain widely. In other words, because collective action problems are both common and serious, there will be, in most instances of political interaction both within and between states, a demand for rule-based cooperation to emerge.

---

22 We can think of the mechanism of response as being primarily intentional: to some extent, people consciously follow rules because they have insight into the nature of collective action problems and see how rules can help (Keohane 2005; Ostrom 1990; Ostrom 2005). But the emergence of rules is also driven in part by evolutionary processes that are not consciously chosen (Tomasello 2009; Heath 2009a; Bicchieri, Duffy, and Tolle 2004; Axelrod 1984; Gintis et al. 2005).

23 Coleman 1990, chap. 10.
2.2.1 Political conflict

The first condition for a collective action problem is that opponents must have conflicting interests. Conflicts arise when people want different and incompatible things and they are mobilized to achieve their different goals, bringing themselves into opposition with each other. Conflicts are *political* when power comes into the equation: that is, when there is pressure to make binding collective decisions that are enforced through coercion.\(^{24}\)

Interdependence is a source of both conflict and power. Interdependence is most often a result of living in proximity to each other, but not always. In a globalized world, interdependencies reach around the world. All such interdependencies create at least the potential for relations of power, and thus for conflicts of interests to become politicized.\(^{25}\)

The existence of political conflict is, of course, a highly non-restrictive condition. Even the most peaceful, stable societies are rife with conflicts, but those conflicts are, for the most part, managed productively and channelled into non-violent forms of contestation. Political conflict can also be open or suppressed.\(^{26}\) Open conflicts, in which contested claims and opposition are openly engaged in, garner the most attention. But a lack of open conflict cannot necessarily be taken as evidence that no conflict is present, i.e. that relations are purely consensual, that no power is deployed, or that interests are fundamentally in harmony. The lack of open conflict can also be a sign that conflict is present but suppressed through overwhelming domination and power.\(^{27}\)

\(^{24}\) Warren 1999b.

\(^{25}\) Bohman 2007.

\(^{26}\) Warren 1999b.

\(^{27}\) Scott 1985.
2.2.2 The possibility of mutually beneficial cooperation

The second condition for a collective action problem—and thus the demand for shared rules—is that political opponents have some interest in cooperation. This condition qualifies the first one. In politics, actors always have conflicting interests: they want at least some incompatible things. But in order for shared rules to be collectively useful to them, their interests cannot conflict all the way down; the conflict, in other words, cannot be truly zero sum.

I argue that this condition, too, takes in most political interaction and excludes relatively little. While particular aspects of many conflicts might be zero sum when considered in isolation, that is unlikely to be true of interactions between opponents taken as a whole. Thus, for example, a conflict over territorial sovereignty, such as in Northern Ireland, might appear to be zero sum: whatever Irish nationalists gain in the authority of the Republic of Ireland over Northern Ireland, unionists lose, and vice versa. But when we pull back from the specific issue of sovereignty to consider the relationships between the actual people involved in their full complexity, it becomes clear that the conflict is not truly zero sum. There are ways for unionists and nationalists in Northern Ireland to manage their relationships with each other (of which the sovereignty conflict is only one part, albeit a large one) that benefit both sides and, alternatively, ways of relating—based on violence, intimidation, and repression—that harm both.28

There are two broad theoretical reasons to expect that most cases of political conflict involve common interests as well as opposing interests. One is the basic human condition of vulnerability. As Hobbes pointed out, “for as to strength of body, the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others that are

28 Even in war, combatants generally have some common interests and work out limited forms of cooperation on that basis. Staniland 2012.
in the same danger with himself.” And our vulnerability to each other is not only physical. As Axel Honneth and others have argued, human beings are fundamentally social creatures, with needs for social esteem and recognition that make us dependent on others, and thus vulnerable to others, in psychic, as well as physical, ways. Human beings in contact with one another nearly always have the ability to do one another considerable harm, and that fact alone means that there is nearly always scope for the development of mutually beneficial systems of rules that prohibit certain harmful actions.

A second reason to expect that in most political conflicts and interactions there are considerable common interests has to do with the complex interdependence entailed by modern economic production. As Adam Smith argued, most goods can be produced with vastly more efficiency when large numbers of people divide their labour and specialize. Many goods can only be produced through the coordinated actions of many people: nobody can manufacture a laptop computer, or, for that matter, write a dissertation, on his own. The point is that the immense productive potential of coordinated action means that most people on opposite sides of political conflicts have considerable scope for mutually beneficial cooperation.

Nonetheless, there are some important cases in which this is not true. In some political conflicts and interactions, the parties’ interests, taken as a whole, are zero sum and the scope for productive cooperation is slim or non-existent. This occurs, for example, when one party benefits

30 Honneth 1996.
31 Smith 2003.
32 For a wider view of the different types of efficiency gains possible through cooperation, see Heath 2006.
more from destructive or violent conflict than it would from peace. Armed groups sometimes fit this description, particularly when they do not have a civilian constituency but, to the contrary, prey on civilians. The Lord’s Resistance Army in Uganda, which relies upon violence, kidnapping, and rape to reproduce itself, is an example, as are armed groups that control resources, such as mines or drugs, that would be unavailable to them in peace time. Some international conflicts, too, fail to demonstrate even limited scope for cooperation. For example, given Hitler’s intention to annex vast areas of the USSR as Lebensraum, and to use its inhabitants effectively as slave labour, the conflict between Nazi Germany and the USSR was probably zero sum. In situations like this, opponents do not face the problem of how to cooperate on the basis of shared rules and trust because both cooperation and trust are inappropriate or impossible. 

2.2.3 Incentive problems

The third condition for a collective action problem is the presence of perverse incentives. This condition builds on the first two. A collective action problem occurs when people have both conflicting interests and a common interest in maintaining a cooperative scheme, together with incentives to defect from that scheme— incentives that, if followed, lead to an inferior outcome for all.

There are a number of well-known models of collective action problems. In interactions involving a large number of people (n-player games) the best-known framework is that of public

33 Le Billon et al. 2000.

34 As was demonstrated when Hitler reneged on the Hitler–Stalin Pact.
goods provision. A public good is a jointly supplied, free-access good.\textsuperscript{35} \textit{Jointly supplied} means that it can only exist if all or most of the people who benefit from it contribute to its production or conservation. \textit{Free-access} means that once the good is supplied, everyone benefits, whether he contributed or not. Classic examples of such goods include street lights, clean air, and shared fish stocks or pastures.\textsuperscript{36} Such goods create collective action problems because while everyone benefits if they are produced or conserved, each individual involved has an incentive not to contribute.

In interactions among a small number of people, the Prisoners’ Dilemma is perhaps the best-known way of representing what is essentially an equivalent problem. A Prisoners’ Dilemma is two-player game in which common cooperation is better for both parties than common defection.\textsuperscript{37} But for each party, the very best outcome happens if it “suckers” the other party by defecting when the other cooperates, while the very worst outcome is to be suckered. In such a situation, each person has an incentive to defect, even though the result of common defection is worse for both parties than common cooperation.

While there is a virtually infinite number of forms of collective action problem, they all share the main characteristic that individual actors have incentives not to contribute to an outcome that is good for all. Examples of collective action problems abound in political and

\begin{quotation}
\footnotesize
\textsuperscript{35} Mansbridge 2014. Economics distinguishes between public goods and common resources (or common-pool resources). Both are non-excludable (or “free-access,” in Mansbridge’s terms), but whereas public goods are non-rival (meaning that one person’s use does not diminish anyone else’s, like a lighthouse), common resources are not (if I eat a fish from our common stock, you cannot eat it) (Mankiw 2001, chap. 11). For our purposes, though, this difference between public goods and common resources is less important than the fact that both are jointly supplied—in the case of public goods, by everyone contributing, and in the case of common resources, by everyone refraining from depleting them. Because of this joint supply, both are subject to similar collective action problems, and on that basis, I refer to both broadly as “public goods.”

\textsuperscript{36} G. Hardin 1968; Ostrom 1990.

\textsuperscript{37} For the story about the prisoners, see e.g. Mankiw 2001, 359–360.
\end{quotation}
social interactions. Violence, or even the possibility of violence, creates stark collective action problems. In many violent conflicts, while both sides benefit from the cessation of costly violence, each side would benefit even more by being able to use or threaten violence while the other side does not. Peace and the common restraint of violence is, in effect, a public good. In weak states or transitional democracies like Egypt, the creation and maintenance of non-partisan bureaucratic institutions often presents collective action problems. While all political parties, or at least the broad constituencies they represent, would benefit from the creation of a strong and non-partisan civil service, each party, when it is in power, has incentives to capture institutions and turn them to partisan advantage. In international politics, arms spending among allies presents a collective action problem. Among NATO countries, for example, each state benefits from the collective defense guaranteed by the North Atlantic Treaty, but each also has an incentive to minimize its own defense contributions while free-riding off the military spending of others.

Incentive problems are present in the vast majority of political conflicts. There are, however, some cases that are ruled out by this criteria. In particular, I have in mind cases that present pure coordination problems—a form of collective action problem that does not create incentives to defect from a scheme of mutually beneficial cooperation. The classic example of a coordination problem is which side of the road to drive on: once a convention to drive on the right is established, nobody has an incentive to defect from the rule by driving on the left.

In that example, there is no underlying conflict of interest: nobody has a stake in whether the rule is to drive on the left or the right. But pure coordination problems can also occur where there are conflicting interests. For example, the technical protocols for communication over the

global Internet have largely been determined by the United States, even though some other states would prefer other standards.\textsuperscript{39} But once those technical standards are in place, no country has an incentive to unilaterally depart from them because to do so would simply cause chaos.\textsuperscript{40} The problem in a coordination problem comes not from incentives but from the existence of multiple equilibria, and thus of not knowing which convention the other player will follow.

The term “collective action problem” is often used as an umbrella term that includes both coordination games and games that, like the Prisoners’ Dilemma, involve incentive problems.\textsuperscript{41} In this dissertation, however, I will use “collective action problem” to refer specifically to situations in which cooperation is threatened by incentive problems—and, by implication, if the problem facing actors is purely one of coordination, it is not, on this account, a collective action problem. The reason for this terminological choice is my belief that incentives to defect, and not coordination, are the main concern of real political actors who wish to cooperate in relatively anarchical circumstances. In deciding whether to cooperate with an opponent, I would suggest that the thing that most often keeps a decision-maker up at night is either the fear of being suckered by a wily opponent or that an opponent might defect for fear of her own treachery (the incentive problem), and not the possibility that she and her well-meaning opponent might get their signals crossed about what cooperative plan they are following (a coordination problem).\textsuperscript{42}

\textsuperscript{39} “A Plaything of Powerful Nations” 2011.

\textsuperscript{40} Cf. Martin 1992, 775–77.

\textsuperscript{41} For example, in Lisa Martin’s (1992) influential scheme, Prisoners’ Dilemmas and the like are “collaboration games,” which, together with “coordination games” and two other game types, are comprised in the general category of collective action problem.

\textsuperscript{42} See also Fearon’s (1995, 385, 401) similar skepticism about whether the “security dilemma” (Jervis 1978; Jervis 1976) and misperception alone are enough to cause escalation in arms races or even war. Rather, it is the commitment problem—that is, the fact that a state may have an incentive to attack another if it anticipates its relative power weakening—that more convincingly explains war.
When it comes to collective action in anarchy, incentive problems—meaning opportunistic defection and the fear of opportunistic defection—are the main challenge, and that is the challenge I want to capture in the term “collective action problem.”

Admittedly, the claim that incentive problems are widespread, and concomitantly that pure coordination games are relatively rare, is somewhat controversial.\(^{43}\) The business of reconstructing the incentive structures facing political actors in order to identify the underlying “game” they are playing is difficult. There are often many possible stories we could tell. In particular, when analyzing a situation that looks, at first blush, like a Prisoners’ Dilemma or public goods problem, and thus includes incentives to defect, it is often possible to construct the game differently so that those incentives disappear. For one thing, we might reassess the preferences of one or more players. If the players in a Prisoners’ Dilemma have, for example, internalized a normative commitment to cooperation such that they now prefer mutual cooperation to suckering the other, the incentive to defect disappears.\(^{44}\) Alternatively, we might identify a new cost to defection that was previously unaccounted for. For example, the game might, after all, be a repeated Prisoners’ Dilemma rather than a one-shot interaction.\(^{45}\) Now the threat of punishment in future rounds makes defection in any given round too costly, and the incentive to defect disappears again. Given the many different ways that any concrete political situation might be interpreted, it can be difficult to tell whether it involves significant incentives to defect or not. And it is more difficult still to generalize about the prevalence of such incentives in political interactions generally, as I have just done in suggesting that coordination problems,

\(^{43}\) The “relative gains” debate in International Relations turned, to a considerable degree, on this question. Baldwin 1993.

\(^{44}\) Bicchieri 2006, chap. 1.

\(^{45}\) Axelrod 1984; Keohane 2005, chap. 5; Mailath and Samuelson 2006.
involving no incentive to defect from cooperative schemes, are relatively rare in political conflict.

Nonetheless, it seems clear that in a great many political interactions, incentives of the sort that are represented in Prisoners’ Dilemmas and public goods provision remain a problem. As I suggested above, the question is whether the actors involved experience the situation as one in which they are tempted to defect or fear that others might be tempted to defect. It may be that the situation actually is a Prisoners’ Dilemma; or it may be that they mistakenly think it is a Prisoners’ Dilemma; or even that there is significant chance that it might be a Prisoners’ Dilemma.\textsuperscript{46} Whatever the case, as long as the possibility of opportunistic defection is on the minds of the actors involved, collective action problems, and how to navigate them, remain a salient part of the situation as it is experienced by those actors.

One benefit of framing our inquiry in terms of the emergence of shared rules and commitments, rather than the more generic terms cooperation or collective action, is that this framework reflects the prevalence and seriousness of incentive problems in political life. Cooperation and collective action are sometimes used to refer to situations that present straightforward coordination problems, in which people’s preferences are nicely aligned, and thus collective action doesn’t face much of an obstacle. To talk of rules and commitments avoids suggesting this kind of easy or harmonious coordination. The language of rules and commitments more easily connotes the need for self-restraint,\textsuperscript{47} for curbing impulses one might otherwise have followed; rules and commitments, in other words, have the connotation of restraining or “biting” in a way that “cooperation” doesn’t necessarily have. That notion of

\textsuperscript{46} In other words, incentive problems can be salient even in games that aren’t Prisoners’ Dilemmas if players are uncertain about others’ preferences or the incentives they face. Jervis 1978.

\textsuperscript{47} See n. 8 above.
restraint is an important part of the experience of rule-making and rule-following. If we conceive of cooperation simply as coordination we make the problem too easy; we fail to recognize why political opponents find it useful to make rules, and also what makes rule-making so difficult, particularly under anarchical conditions.

In summary, collective action problems provide valuable insight into the nature of rules. To approach rules through the problems they solve is, as I suggested above, to give a functionalist account of them. On this view, people make rules precisely to escape from patterns of self-defeating action in situations where individual rationality prescribes strategies that leave everybody worse off. Rules, in other words, are tools. By thinking of rules this way—that is, thinking of the functions they serve for those who create and maintain them—we gain insight into the rules themselves and into when and how political opponents might succeed in creating them.

2.2.4 Thinking of rules as tools: some caveats

Before moving on, though, we should recognize some important limitations, even dangers, to thinking about the emergence of rules in functionalist terms. Identifying those limitations is not difficult: functionalist thinking has come in for a lot of criticism over the years.⁴⁸ There are at least five limitations and dangers to be aware of.

First, the fact that rules would be useful does not mean that people will create them. One problem is that the creation of institutions for the management of collective action problems is

---

itself subject to “second order” collective action problems. In order to succeed in building mutually beneficial systems of rules and commitments, political opponents have to substantially overcome, transform, or circumvent those perverse incentives. There is no guarantee that they will succeed. Collective action problems, in effect, create a demand for rules that help to realize common interests. Whether, and how, that demand is met is another question.

Second, thinking of rules as solutions to collective action problems can lead to an overemphasis on the productive as opposed to the distributive dimension of institutional bargaining. Rules can help political opponents realize common interests—that, from this functionalist point of view, is why people make and follow those rules. But it is important to keep in mind that opponents have opposing interests as well as common ones. They aim not only at interests held in common (“making the pie bigger”) but also at interests that are incompatible (“taking more of the pie”).

This is an important point to make because from a normative point of view, “overcoming collective action problems” is by no means the only concern we should have. Systems of rules can be mutually beneficial while still being grossly inequitable and unjust. In economic terms, a system of rules might represent a Pareto-improvement (it makes at least one player better off and nobody worse off) or even Pareto-optimality (no alternative scheme would make anyone better off without making someone worse off), but that says nothing about where it lies on the “Pareto frontier”—that is, about how evenly the gains are distributed.

49 Elster 2005.
51 Lax and Sebenius 1986.
Further, emphasizing productive over distributive dimensions of bargaining can also mean overlooking relations of power and domination. Talk of common interests and mutually beneficial exchange or production is sometimes taken to imply that cooperation is fully consensual, or that no coercion is involved—an assumption sometimes made about market exchanges. In many cases, however, that assumption is mistaken. Possible rules or deals differ in how they distribute benefits between the people who make them, and those people will generally bring whatever power and resources they can to bear on that distributional struggle. Since rules are shaped by power, and power is usually unevenly distributed, we should expect the distribution of goods created by systems of rules to usually be unequal.\textsuperscript{52} The tendency to overlook conflict and power is a recurrent criticism of functionalism.\textsuperscript{53}

Third, the common goods secured through rule-based cooperation are not necessarily good in a normative sense. Criminals and despots, too, can create rules to overcome collective action problems among themselves, which only makes them better at preying on others.\textsuperscript{54} The assumption in this study is that the emergence of shared rules in anarchical political conflicts is usually a good thing. But in any particular case, that judgment is conditional on the shared purposes (or “functions”) that the actors are seeking to make rules serve. In the case of Northern Ireland, for example, those common goods were indeed, for the most part, good. The regime that emerged secures some undoubtedly good things, including a reduction in violence, greater economic prosperity, mutual recognition, and greater political and social equality between the

\textsuperscript{52} Knight and Johnson 2011.
\textsuperscript{53} Mills 1959.
\textsuperscript{54} Odell and Tingley 2013, 145; Warren 2008.
two communities. But whether the shared goods that a particular set of rules secures are good in a normative sense is always an open question.

Fourth, to think of rules as tools is to assume that they are purposive constructions in the first place. In reality, though, our actions are always influenced by dispositions, habits, and cultural assumptions that we do not choose, and may not even know about. Conformism, whether conscious or unconscious, can result in the creation of rules and norms that, paradoxically, leave everyone worse off, not better.55

Fifth, and finally, thinking about rules instrumentally does not imply that people’s purposes are only self-interested or materialistic. The recent “systems” turn in democratic theory is a case in point.56 Democratic systems theory adopts a roughly functionalist approach to the study of democratic institutions, but the goods or “functions” are democratic ones, such as inclusivity, equality, and autonomy.57 Likewise, some of the key goods secured in the Northern Irish settlement are ethical or normative values, including equal recognition.58 Thus, the “rational choice talk” involved in thinking of rules as responses to collective action problems does not imply that all goods can be defined in terms of material self-interest. (But, of course, rational choice theory itself doesn’t imply that either.)59

There are, thus, some important caveats to keep in mind when thinking of rule-making as a response to collective action problems. Nevertheless, despite these limitations, understanding collective action problems is an important starting point for understanding why and how rules

55 Bicchieri 2006, chap. 5; Elster 1989b, chap. 7.
56 Parkinson and Mansbridge 2012.
57 Warren 2012b; Dryzek 2009.
and joint commitments emerge in the midst of conflicts. Making and following rules together is an activity, it’s something people do, and collective action problems are a salient feature of the situations in which people choose to engage in this activity.

### 2.3 Difficult circumstances for rule-making

I now turn to the second part of the dissertation problem: the challenges raised to building shared rules under relatively anarchical conditions. I use the term “anarchy” to refer to a complex of two conditions: weak governing institutions and thin social fabrics. In such circumstances, political actors who want to establish new rules and institutions must, in effect, “rebuild the ship at sea”—they must generate resources for collective action as they go along. In this section I will examine in more detail the challenge of building shared rules in anarchy.

#### 2.3.1 Weak governing institutions

The first condition I refer to with the term “anarchy” is weak (or missing) governing institutions. Weak states, divided societies, civil conflicts, and international politics all, in different ways and to different degrees, exhibit this characteristic—in comparison, at least, to politics within the boundaries of strong and legitimate states.

The weakness of governing institutions shows up in both a lack of authority and capacity to make and enforce binding decisions. As to authority, an anarchical context, in this sense, is one in which few, if any, rules or institutions are commonly recognized as authoritative by the

---

60 As I explained in Chapter 1, while my use of the term “anarchy” is broadly consonant with the way the term is used in International Relations theory (e.g. Bull 1977; Waltz 1979), my primary purpose is not to characterize international relations per se. Rather, I am interested in these two conditions of weak institutions and thin normative consensus wherever they occur.
people whose interactions we are studying. As a matter of degree, a context of interaction is more anarchical if there are fewer rules, if the rules that exist are trivial or unimportant and do not impinge on people’s actions, or if rules are considered optional or aspirational as opposed to binding.

The lack of authoritative governing institutions is perhaps most obvious in violent civil conflicts involving secession claims, since at least one group rejects the legitimacy of the state. To organize violent capacities to be used against the state is necessarily to contest its authority. In weak states or transitional democracies, such as Egypt, a similar situation prevails. While such states might not be subject to secessionist claims against the legitimacy of the state as a territorial unit, they are subject to contestation of the legitimacy of state institutions to make binding rules within that territory. And even where citizens recognize the formal authority of the state to make binding rules, they might as a matter of practice consider compliance with those rules to be optional—something that can frequently be seen, for example, in tax compliance rates.\(^6\) Thus, civil wars and weak states generally lack legitimate coercive authority.

In international politics, the lack of legitimate authority might be less obvious, given the existence of international law, international organizations, and other sources of authority. Nevertheless, international and transnational politics remains significantly anarchical in comparison to politics within the state. While there are many rules, what is lacking is the authority to *enforce* those rules. While there is authority vested in international law and international organizations, such authority is mostly limited to voluntary compliance and consent, much more so than in domestic politics. At bottom, the normative basis of international

\(^6\) Rothstein 2005a, chap. 1.
law remains state sovereignty. As Kenneth Waltz puts it, “none is entitled to command; none is required to obey.”62

Furthermore, the legitimate coercion that does exist in international law remains highly circumscribed. Sovereign states under international law are far more legally autonomous than individuals in modern states. While states are subject to legitimate coercion with respect to only a very small number of fundamental obligations—primarily those having to do with international peace and security, which the United Nations Security Council is entitled to enforce63—the same cannot be said of me as a Canadian citizen vis-à-vis the Canadian state. Canadian criminal law sets out myriad rules constraining my actions, none of which depends upon my consent for its enforcement. Thus, the growing authority and legitimacy of international organizations might lessen, but it does not eliminate, the anarchical character of international politics, because although there are rules, they still rest on the voluntary observance and consent of states.64 What is missing in international politics, in contrast to politics within the modern state, is institutions with the authority to coerce.

In addition to lacking authority, governing institutions can also lack the capability to effectively enforce rules. Violent conflicts again provide the most obvious cases of such institutional weakness. In civil conflicts and weak states, state institutions have, by definition, diminished coercive capacities. The fact that an insurgency, for example, is capable of putting up sustained resistance to the state demonstrates that the state’s writ is weak, at least in some parts of the country. The amount of overt coercion in weak states—the police patrols, the armed

62 Waltz 1979, 88; see also Jackson 2000.
checkpoints, etc.—is evidence of their coercive weakness, not their strength, vis-à-vis their citizens.

International politics, too, is marked not only by a lack of coercive authority, but also by a lack of coercive capacity. The ability of states—including the most powerful ones—to impose compliance with international law on other states is relatively weak in comparison to the power of most states over their citizens. To be sure, the coercive weakness of international institutions is a matter of degree. There are occasional cases of effective enforcement: the 1991 Gulf War was an effective sanction of Iraq’s invasion of Kuwait, and the bombing campaigns in Kosovo in 1999 and in Libya in 2012 were reasonably effective enforcements of human rights norms. But the Russian invasion of Georgia and, more recently, Crimea, ethnic cleansing and killing in Sudan, and the development of nuclear weapons in North Korea are examples of recent violations of international laws and norms that were not deterred.

2.3.2 Thin social norms

The second characteristic of relatively anarchical situations is the relative thinness of social norms and shared cultures. In international and global politics, sheer cultural diversity thins out the stock of norms that can been taken for granted among political opponents, and that might otherwise serve as important resources for managing conflict. The same is true for politics in deeply divided societies and civil conflicts. These are situations in which salient identities and social norms (including norms of group-based retribution or ethnic loyalty) often exacerbate conflict rather than serve as a resource for containing it. In any particular anarchical context, the

lack of normative consensus is a matter of degree. In international politics, states do, in fact, share a considerable stock of taken-for-granted norms and assumed identities. And in the context of divided societies or civil war, there is often considerable cultural and normative agreement underlying the strident disagreement. Nonetheless, to the extent that a given situation is anarchical, as I have defined that here, it will exhibit a relatively thin fabric of shared norms.

2.3.3 Rebuilding the ship at sea

What might be the effect of such anarchical conditions on the emergence and maintenance of shared rules? The main effect, I argue, will be to make rule- or institution-building harder. The reason is that institutions have “self-thickening” tendencies. The prior existence of rules offers resources, normative and coercive, for making new rules. And, mirroring that, establishing new rules is particularly difficult in relatively anarchical circumstances where those resources are missing.

How do rules and institutions facilitate the making of new rules? They do so in at least two ways. First, the existence of prior institutions means that interactions between political opponents already have a significant normative structure. Where norms are already in place, it is much easier to expand those obligations than to establish such obligations for the first time. This has to do, first, with the content of rules. The rules that exist provide a store of agreed-upon principles, understandings, and procedures from which new rules can be built. But it also has to

---


67 On the “narcissism of minor difference” in ethnic conflict, see Ignatieff 1998.

do with the fact that the prior existence of rules implies a relationship of mutual accountability that does not exist in more purely anarchical situations. One way of putting it is this. Where opponents are adding a new rule or commitment to an already-extensive structure of rules, they have to bring each other, one way or another, to endorse the content of that new rule. But where they are attempting to establish mutually binding rules for the first time, they have to achieve agreement not only on the content of those rules, but also on the notion of being bound by rules itself. Establishing that notion of mutual accountability per se is a significant normative accomplishment in anarchical circumstances. In the worst circumstances of war and violence, the idea that one owes anything to an enemy cannot be taken for granted.

The second way institutions facilitate the making of new rules is through their coercive capacity. Strong institutions, like those of liberal-democratic states, make rule-making safe. Within such states, for example, legislators do not have to question whether rules that are decided on will be accepted and followed by those who lose a legislative vote as well as those who win. The contest over whether to enact a new rule or law can, of course, be a fierce one. But the difficulty of coming to agreement is primarily about conflict over the content of the law and the consequences it will have, rather than fear that it will not be upheld once enacted. In relatively anarchical contexts, by contrast, fears of opportunistic defection pose a much larger challenge to establishing shared rules and joint commitments. Where political opponents know that enforcement mechanisms are weak, they will be reluctant to make commitments for fear that keeping those commitments would open them to being suckered by less scrupulous opponents.69

To return to the metaphor introduced in Chapter 1, the challenge of establishing shared rules and joint commitments under such conditions is akin to rebuilding a ship at sea. That image

69 Fearon 1998.
captures something of the paradox of how new rules and commitments emerge in otherwise rule-thin environments. Ship-building is normally done on land because the ship cannot be expected to float until it is largely complete. To rebuild a ship at sea means, paradoxically, to require the ship to float even while it is being built. The problems of ship-building have to be solved not in the comfort and safety of the dry-dock, but rather when the voyage is underway. Rebuilding a ship at sea is a metaphor for undertaking an activity even before the seeming prerequisites for that activity are in place; it means generating the resources you need as you go along.

In pessimistic moments, we might think that establishing rule-based cooperation is like ship-building: it has to be done “on land.” That is, cooperation requires the institutional supports for cooperation to already be in place. How can political opponents trust each other enough to make and keep agreements without such supports, given the dangers and uncertainties of doing so? And if they can’t make and keep agreements, how can they ever build the institutions they need in order to cooperate? There seems to be no way to break out of mutually harmful patterns of interaction.

But we know from observation that such vicious circles are not always inevitable or irreversible. Political opponents do sometimes succeed in building, step by step, new political realities—new forms of cooperation, together with the institutional arrangements, both normative and coercive, to support that cooperation. When they do so, they succeed, in some sense, in getting something from nothing, in generating new resources for cooperation that previously did not exist.

__________________________

70 Waltz 1979.
2.4 Conclusion

Political opponents in anarchical circumstances often face a difficult challenge. Escaping from mutually harmful conflicts requires that they establish mutually beneficial understandings, agreements, or rules. But in anarchical circumstances, that is no easy task. One the one hand, they lack the prior normative commitments and shared meanings that limit conflict and facilitate agreements in stable societies. On the other hand, they also lack the safeguards that strong institutions provide through their ability to enforce rules. As a result, success depends upon their ability to “rebuild the ship at sea”: to generate resource for collective action and joint commitment as they go along.
The previous chapter presented the problem this dissertation intends to address: how do political actors build new shared rules and institutions in relatively anarchical circumstances? How do they rebuild the ship at sea? In this chapter, I will begin to lay the foundations of the answer I will develop throughout this dissertation: a theory of how deliberation contributes to rule-based cooperation in relatively anarchical circumstances. In many ways, deliberation is a promising response to conflict in difficult circumstances. In deliberation, participants offer reasons and justifications to each other. They attempt to persuade each other, to find common standards of justification, and to come to a shared understanding of their common problems and how they should be addressed. If political actors succeed in that endeavour, they are clearly taking significant steps toward building and maintaining cooperative rules and institutions.

But despite the evident promise of deliberative responses to conflict, our theories of how deliberation might have such effects are still lacking. The problem is that despite the potential benefits of successful deliberation, it can also be difficult, dangerous, and fragile.¹ Those who deliberate sincerely—who speak openly about their beliefs and intentions, stand by their commitments, and count on others to do the same—can often be taken advantage of by others. In order to take the risks and trouble associated with deliberation, people have to feel reasonably secure and confident that others will play by the rules. In strong, liberal-democratic states, those securities are provided, on the one hand, by state-backed institutions and law,² and on the other

---

¹ Warren 2007, 276–77. Habermas (1996, 147) speaks of good reasons as providing only a “weakly motivating force.”

² Habermas 1996, 28–41; Mansbridge 2001; Mansbridge et al. 2010, 83.
hand, by relatively cohesive civic cultures and background normative consensus. But those, of course, are precisely what is missing (or relatively weak) in the difficult circumstances this dissertation is concerned with. Theorists of deliberation have not yet responded to this challenges as well as they need to if we are to understand how deliberation might get a grip in these difficult circumstances.

In this chapter, I will set out this challenge in detail and then present a preliminary account of the answer I will develop in the remainder of the dissertation. I will argue that the key to understanding deliberation in relatively anarchical circumstances is to better theorize the connections between deliberation and *relationships of mutual accountability*. The “forceless force of the better argument” comes in part from the social commitments implied in the act of addressing those arguments to one another. Reasons (specifically, public reasons) have their characteristic force only within the context of a group of actors who take themselves to be accountable to each other in terms of reasons. Relationships of mutual accountability create the “space of reasons” within which argument and persuasion become reliable ways of coordinating action.

On the one hand, that connection between deliberation and mutual accountability helps to explain what makes deliberation so difficult in relatively anarchical circumstances. Where state-based securities and institutions are weak, people have to be all the more confident in each

---


5 Habermas 1999, 450.

6 Brandom 1994, 5; Baynes 2007, 72; Erman 2010.
other’s commitments if they are to depend on deliberation to manage conflict—and the burden this places on trust is often too great.

But on the other hand, the connection between deliberation and mutual accountability also suggests how deliberation might overcome that challenge. Deliberation not only depends on relationships of mutual accountability; it also helps to generate them.\textsuperscript{7} Deliberation can be a way for political opponents to build and demonstrate the very commitments and accountabilities required for further deliberation and for a deliberative politics that supports cooperation on the basis of shared rules. The result can be (though it need not be) a virtuous circle, one that allows deliberative politics to get a foothold even under difficult conditions. With time, deliberative capacity based on interpersonal commitments may then feed into institutions and social consensus that help to stabilize productive politics over the long run.

The chapter begins by building up some conceptual preliminaries, including what I mean by “deliberation.” While there is broad agreement that deliberation is about the give and take of reasons, confusion remains about what exactly counts as a “reason.” In section 3.1, I will give an answer to that question, and explain how deliberation relates to rhetoric and “emotional” speech, on the one hand, and bargaining, on the other.

Section 3.2 then elaborates the challenge that relatively anarchical circumstances present to deliberation and outlines the solution I intend to develop. In the absence of strong institutions and social norms, deliberation relies more heavily on interpersonal trust and mutual accountability. But deliberation may also be able to build trust and accountability, and in doing so contribute to the task (outlined in the previous chapter) of “rebuilding the ship at sea.”

\textsuperscript{7} Habermas 1998, 63–64; Arendt 1998, chap. 5; Warren 2006.
Finally, in section 3.3, I address an outstanding conceptual problem for deliberative theory. Political interactions under anarchical conditions tend to be suffused with coercion and other forms of non-deliberative influence that threaten to displace deliberation. I argue that deliberation can still be effective even in such circumstances, and that properly arranged, symmetrical uses of coercion can even support deliberative influence.

3.1 What is deliberation?

Despite the fact that deliberative democratic theory has come to occupy a central place in political theory, there remains considerable diversity of opinion about what exactly deliberation is. Nonetheless, there is broad consensus on the conceptual core of deliberation. Deliberation is best understood as the form of influence generated by the give and take of reasons, or, in other words, by mutual justification.

Conceptually, then, deliberation is relatively straightforward, at least in its core meaning of “influence through reason-giving.” Distinguishing deliberation from other kinds of talk-based influence, however, is trickier. Saying what counts as a “reason” and what does not has been one of the more difficult challenges deliberative theorists have faced. Is reasoning limited to formal or logical argument, or can other forms of communication, including rhetoric and story-telling, count as “reasons” as well? Can appeals to self-interest count as reasons, or only appeals to the common good or universal moral principles? What about non-verbal, symbolic displays of emotion, solidarity, or protest? If deliberation is to be a useful concept, we have to be able to

8 Neblo 2007; Bächtiger et al. 2010

9 Habermas 1996. Warren (2002) discusses deliberative influence as a medium of social coordination analogous to other media such as money and coercive power.
answer such questions. Not all talk can be called deliberative, as otherwise the concept would lose its distinctiveness, along with its analytical and normative value.\textsuperscript{10}

When it comes to distinguishing deliberation from other forms of talk, there are two important points to keep in mind.

3.1.1 Deliberation and emotion

The first point is that “reason” should not be interpreted as the opposite of emotion or affect. It is sometimes argued that the idea of deliberation promotes a concept of the self that excludes, or devalues, the non-rational aspects of human experience. To focus on reason, in other words, is to exclude the emotional, embodied, concrete self.\textsuperscript{11}

If that were true, it would be a troubling problem for deliberation. To begin, it would imply that the idea of deliberation rests upon a conception of the human self that is simply false. We are not dispassionate, detached, unencumbered “rational” selves.\textsuperscript{12} To the contrary, we are all rooted in societies, communities, and personal relationships, and those concrete ties affect us and inform our thoughts and intentions. Cognitively, we are subject to all manner of biases, frames, and emotions that we do not choose and are not even aware of.\textsuperscript{13} Motivationally, we are driven as much by passions—by desire, fear, and affection—as by “cooler” or more calculated considerations.\textsuperscript{14}

\textsuperscript{10} Bächtiger et al. 2010.


\textsuperscript{12} Sandel 1984.

\textsuperscript{13} Ksiazkiewicz and Hedrick 2013; Kahneman 2011; Heath 2014b.

\textsuperscript{14} Cf. Hume 1969.
To overlook the emotional, embodied, social self—if that is what deliberation does—would be court a number of dangers. First, it would mean that legitimate emotional appeals, expressions of sympathy and solidarity, and other forms of speech that are not overtly “rational” or analytic might be unfairly discounted in decision-making. Second, to adopt such a narrow view of the self might be to overlook forms of domination that operate on a sub-conscious level. To the extent that political elites can influence people by subtly framing issues, triggering biases, and manipulating fears, the idea that they form their opinions through the “force of the better argument” might be naïve. Finally, by emphasizing the “rational” over the emotional, deliberative theory might cut itself off from understanding the positive potential of speech to foster empathy and emotional solidarity.

These are important concerns. But as criticisms of deliberation, they are largely misdirected because deliberation—and the concept of rationality it implies—is not exclusionary in this way. “Reason” is best understood as a formal concept. It points to the sense of conviction or endorsement that good reasons carry, and not to their particular content. “Reason” refers simply to any consideration, desire, or inclination that an actor takes to be a reason to act—including inclinations that come from emotional connections, dispositions, and solidarities. (I will have more to say about the concept of practical reason, and the moral psychology it implies, in Chapter 5.)

That implies that when it comes to distinguishing deliberation from other forms of talk-based influence, we cannot interpret “reasons” in a restrictive way. A reason is simply a valid

---

15 Iris Marion Young (2000) calls this danger “internal exclusion.”

16 Morrell 2010.

17 Korsgaard 2008a; Kant 1997, 27.
justification, an answer to a “why” question that a person endorses as, in one sense or another, right or compelling.18 No pre-defined type of consideration can be ruled out as reasons, and that goes for motivations associated with passions and with the embodied, socially connected self. While early formulations of deliberative theory tended to emphasize formal argumentation, most theorists now recognize that a wide range of forms of communication—including rhetoric, story-telling, expressions of emotion, and symbolic protest—can also serve as reasons.19 Emotional expressions, stories, and other forms of less formally argumentative communication often have an implicit justificatory purpose.20 In this, they have the same potential as more formal argumentation to bring about rationally motivated agreement.

In fact, emotionally expressive stories can play a key role in deliberation. Jürg Steiner relates the impact of a story told during somewhat tense deliberations between former members of the FARC and former right-wing paramilitaries in Colombia.21 This story turned out to be a transformative moment in the conversation, after which participants began to engage with each other’s statements seriously rather than talking past each other. The speaker told the story of how he had come to join the FARC by being influenced as a teenager by a friendly recruiter, by being given opportunities to study, and by the contrast to the cruel treatment he received while in the army. While there was no explicit argument in the story—indeed, when another participant pressed the story-teller to relate his story to the topic of conversation, he was unable to do so coherently—it arguably served as a reason in the deliberations. Those who listened found the

18 Mansbridge 2007, 261.
19 E.g. I. M. Young 2000; Dryzek 2000; Chambers 2009a; Neblo 2007; Chambers 2014.
20 See also Chambers 2012, 59: “It is not the form that determines whether something is a reason but rather the function or purpose it serves in a discourse.”
21 Steiner 2012, 72–74; see also Risse 2000, 15.
story compelling in the same way arguments can be compelling. Even those who were opposed to the FARC found themselves compelled to recognize the speaker as being fundamentally like them, as someone who was well-intentioned but pressed by difficult circumstances.

If the concept of reason is formal, as I’ve argued here, it might sound as if anything goes—any influence through talk might count as deliberative. But it isn’t so. Some forms of talk-based influence are excluded from deliberation. The point, however, is that those exclusions cannot be defined in advance by their content. Rather, they are defined by the fact that they lack the reflective endorsement of the actor being influenced. That actor does not take them up as reasons for action. That is what’s problematic about framing, manipulation, and implicit bias. Ways of influencing people that bypass or undermine their ability to actively take up and judge the grounds of their choices are non-deliberative. What excludes manipulative or implicit forms of influence from being reasons is not that they have emotional content, but rather that the actor does not endorse these inclinations as good ones—and, indeed, he or she is usually unaware of them. Thus, as Michael Neblo puts it (glossing Habermas), “the opposite of reason…is not emotion, but rather unlegitimated power.”

Deliberation, then, should not be equated with dispassionate objectivity or formal argument (though it can take these forms). The hallmark of deliberation, rather, is that actors influence each other in ways that they can recognize and endorse as binding or compelling. Statements that carry this rationally binding force can take a wide range of forms.

---

22 Some of the confusion about what counts as a “reason” comes, I think, from mixing the formal concept of a reason with the substantive evaluation of whether a reason is good. This happens in everyday language: we say “that’s not a reason to do x” when what we mean is “that’s not a good reason to do x.” Reasons we don’t agree with can still be reasons: they are reasons if someone reflectively endorses them as reasons for his or her own actions or beliefs. What makes them reasons is the role they play in that person’s choices; what makes them good reasons (if they are good reasons) is another matter. See Brandom 2000, 183.

23 Neblo 2007, 531.
3.1.2 Deliberation and bargaining

The second point to keep in mind when distinguishing deliberation as a form of communication is that while deliberation necessarily excludes coercive bargaining, it does not exclude reasons of self-interest and non-coercive forms of bargaining and negotiation. Early, “classic” formulations of deliberative democratic theory tended to conceive of deliberation in explicit contrast to self-interest. Deliberation, on this view, is defined by its rationality, and rational arguments appeal to universal moral principles and the common good, as opposed to arbitrary will and self-interest. Deliberation about the common good was contrasted to the aggregation of individual preferences through voting. In deliberation, self-interested claims were assumed to be ruled out, either through the moral restraint exercised by participants in deliberation, or through the “civilizing force of hypocrisy”—the constraints imposed by the fact that appeals to one’s own interest would simply fail to move other participants, or would even incur social sanctions. In the International Relations literature, a similar opposition between deliberation and self-interest is assumed in the distinction between “arguing” and “bargaining.”

But in recent years, the place of self-interest in deliberative theory has undergone significant revision. “Reasons” cannot be the opposite of “interests” because interests, including appeals to self-interest, can themselves be valid reasons. People can and should advocate their own interests in deliberation. Good political decisions rightly reflect the interests

26 Elster 1998.
28 Mansbridge et al. 2010.
of those affected by them, and, generally speaking, those interests are best known and articulated
by the people whose interests they are. Where there is a common good that a political
arrangement or accord might embody, the frank articulation and advocacy of interests helps to
clarify what that good is. And where interests conflict, the articulation of those interests clarifies
the conflict as well—including conflicts that might be overlooked or papered over if less
powerful groups were not permitted to vigorously defend their own interests.29

That is not to say that there are no constraints on advocating one’s own interests in
deliberation. As the classic theory suggests, there are both normative and practical limits to
appeals to self-interest. Normatively, the idea of deliberation embodies standards of equality and
mutual respect, and those requirements militate against appeals to self-interest that are
exploitative or unjust. Thus, as Mansbridge et al. argue, “the desire to become wealthier come
what may” is ruled out as a justification, but “the desire to become as wealthy as possible
consistent with a fair level that others (i.e. equal citizens) find acceptable” is not.30 Furthermore,
there are practical limits set by the need for appeals to self-interest to function as justifications to
others, since others will often find such appeals unconvincing. The point, however, is that neither
normative nor practical considerations rule out self-interested justifications per se, only certain
instances of them. So long as appeals to self-interest are acceptable as reasons, they are a part of
deliberation.

Admittedly, it might seem strange to think that other people’s interests could provide me
with a reason to accept a particular agreement or rule. But it is not so strange. It is simply an
expression of what Rawls called being “reasonable,” which includes treating other people as

29 Mansbridge et al. 2010, 73.
30 Mansbridge et al. 2010, 76.
“self-authenticating sources of valid claims.”31 To treat other people’s interests as reasons for my own actions—within the framework of a fair and non-exploitative scheme of cooperation—is part of what it means to relate to others as free and equal persons.32 This lies entirely within the deliberative ideal.

In distinguishing deliberative from non-deliberative forms of negotiation and bargaining, then, the key issue is not the content of the reasons employed—e.g. whether they appeal to self-interest or the common good—but rather the mode of influence at work. What makes negotiation or bargaining deliberative is that reasons are, in effect, doing the work of bringing about agreement. That is to say, the people involved must be convincing each other that a given outcome or arrangement is, in one way or another, good.

What can “good” mean in the context of negotiations? In some cases, it might mean that the outcome straightforwardly maximizes the utility of everyone involved. If so, the result is a “fully integrative agreement” in which everyone gets everything he or she wants.33 In such negotiations, the interests of the participants are wholly aligned, and deliberation is simply about discovering the best solution for both.34 More commonly, though, such fully integrative agreements are unavailable because participants have at least some conflicting interests in how

31 Rawls 2005, 32. Or, in the more directly Kantian formulation, treating the other as an end in himself. See Korsgaard 2009, 192

32 Note that this is not the same as altruism. In altruism, other people’s interests become my interests, whether through shared identity or emotional attachment. By contrast, I am talking about upholding another’s interest not because it is also mine, but out of respect for the legitimate goals and desires of the other. Cf. Korsgaard 1993.

33 Warren and Mansbridge 2013. See also Mansbridge 2009a and Mansbridge et al. 2010.

34 Such negotiations occur where the underlying collective action problem is one of pure coordination with no underlying conflicts of interest. In that case, negotiation or deliberation can be purely productive (about making the pie bigger) and not at all distributive (slicing up the pie). Fritz Scharpf (1997, chap. 6) refers to such negotiations as “problem-solving.” As I argued in Chapter 1, though, we should expect such circumstances to be quite rare.
the benefits of cooperation are to be divided.\textsuperscript{35} In that case, all possible agreements require one or both participants to give up something they want. Nevertheless, participants can still make appeals to each other on the basis of reasons each other can accept—only in this case, the reasons will have to do with the outcome representing a fair bargain or a principled compromise.\textsuperscript{36}

By contrast, negotiation and bargaining become non-deliberative to the extent that the give and take of reasons is no longer what determines the outcome but rather the exercise of other forms of power. In bargaining, non-deliberative influence generally takes the form of threats (including the threat of walking out of negotiations) or deceit and misrepresentation. Influence wielded through threats differs fundamentally from deliberative influence, because threats take their power not from the cognitively compelling force of a valid claim, but rather from one party’s ability to unilaterally impose harm on the other.

Influence wielded through deceit or misrepresentation is also fundamentally different from the “force of the better argument.” Like the use of threats, deceit and misrepresentation are forms of coercion. Mansbridge et al. provide a good definition of coercive power: “A’s preferences or interests causing B to do (or changing the probability that B will do) what B would not otherwise have done through the threat of sanction or the use of force.”\textsuperscript{37} Deceit is a form of force. Whereas threats affect the victim by engaging his choice, force does not. Force most obviously includes physical force. But taking advantage of an opponent through deceit does essentially the same thing: it is a way of getting someone else to act against his own interests.

\textsuperscript{35} Cf. Warren and Mansbridge 2013, 93. Here I am grouping together, for the purpose of simplicity, what they call “partially integrative agreements” and “fair compromises.”

\textsuperscript{36} P. Jones and O’Flynn 2013; Heath 2001a, chap. 6.

\textsuperscript{37} Mansbridge et al. 2010, 17.
(and in the perpetrator’s interest) without engaging his choice, since, because of the deceit, he
does not really know that he is doing so.38

To say that coercive force, in its various forms, is non-deliberative is not to say that it is
always unjustifiable or even incompatible with deliberation. Most negotiations are probably a
mix of reason-giving and other forms of influence.39 Indeed, in situations in which
misrepresentation, threats, and other tactics of pure bargaining are likely to be deployed,
participants in negotiations may in fact be more likely to approach the deliberative ideal by
balancing coercion with coercion rather than by abstaining from it.40 Bargains that result from
the symmetrical application of power—where participants are roughly equal in power—may end
up approximating a fair, that is, rationally justifiable, compromise.41 Nonetheless, while the use
of threats, misrepresentation, and other purely strategic bargaining tactics may be justifiable from
the point of view of a broadly deliberative ethic, as a matter of analytical clarity, it is important
to recognize that they are not themselves deliberation. I will return to the question of the
relationship between deliberation and coercion later in this chapter.

To summarize, deliberation is the form of influence that takes place through the give and
take of reasons. What marks something as a reason is the claim it makes to being valid or
compelling. Influence through reasons needs to be distinguished from other ways of influencing
one another through talk, but the distinction between deliberation and non-deliberation is not

38 Mansbridge et al. 2010, 18; Warren 2012a, 12–13, citing Bok 1978.
39 Warren and Mansbridge 2013, 94.
40 This is the logic of the “second best.” Though ideally disputes should be solved by arriving at reasoned or
principled compromise, if coercive force is likely to be used, it is better that both parties use it. Mansbridge et al.
2010.
41 On the ideal of equal power in bargaining see Habermas 1996, 108; Knight and Johnson 2007; Warren and
Mansbridge 2013.
drawn according to any pre-defined content, but rather according to whether participants are
moved by their reflective endorsement of claims.

3.2 The problem of anarchy in deliberative theory

With that conception of deliberation in place, I now want to return to the problem I presented in
Chapter 2: the problem of building shared rules in anarchy, or rebuilding the ship at sea.
Relatively anarchical circumstances are in many ways tough cases for deliberation. I argue that
in order to address these problems, deliberative theorists need to make two conceptual moves.
The first is to distinguish the problems of pluralism and disagreement from the problem of trust.
The second is to give an account of the role of supportive institutions and “lifeworlds” in
deliberation and what that implies for deliberation in institutionally thin contexts. Finally, I will
say a few words about what resources deliberative theorists might have to respond to these
challenges.

3.2.1 Deliberation and trust

Efforts to apply deliberative theory to relatively anarchical cases have been gathering steam in
recent years. There are studies of deliberation in divided societies and civil conflicts,46 global

43 Habermas 1996; Warren 2012b.
45 Korsgaard 2009, 190. On the nature of collective intentions and the related notion of group agency, see Gilbert
46 D. Jones 1999; O’Neill 2003; O’Flynn 2006; Drake and McCulloch 2011; Steiner 2012; Luskin et al. 2014.
democracy,\textsuperscript{47} and international relations.\textsuperscript{48} Such efforts often focus on the challenge posed by pluralism and deep-seated disagreement. People on opposite sides of conflicts are often strongly committed to nationalist, ethnic, or religious worldviews that are incomprehensible or repugnant to each other. Successful deliberation requires that participants offer each other convincing reasons, but in the face of deep, and apparently irreconcilable, difference, we might wonder whether any shared reasons are available.

Scholars have made considerable progress in responding to this challenge, in part by clarifying the sorts of agreement that deliberation can bring about.\textsuperscript{49} Deliberation does not have to result in consensus in order to transform conflicts. It can also bring about “incompletely theorized agreements,”\textsuperscript{50} meta-consensus,\textsuperscript{51} principled compromises,\textsuperscript{52} fair bargains,\textsuperscript{53} and other forms of agreement short of full consensus.\textsuperscript{54} While there is no guarantee that in any given case political opponents will be able to reach an agreement, it is also a mistake to suppose that the fact of conflicting values and judgments makes agreement impossible.\textsuperscript{55}

It is important to notice, however, that clarifying the possible content of shared reasons in deliberation does not directly address the problem of trust—that is, the problem of whether

\textsuperscript{47} E.g. Habermas 2001; Dryzek 2006; Bohman 2007; Erman 2013.


\textsuperscript{49} E.g. Bohman 1996; Rawls 2005; Dryzek 2010, chap. 5.

\textsuperscript{50} Sunstein 1995.

\textsuperscript{51} Dryzek and List 2003; List 2007; Dryzek and Niemeyer 2010.

\textsuperscript{52} Gutmann and Thompson 1996; Bohman 1996.

\textsuperscript{53} Rawls 2005.

\textsuperscript{54} Mansbridge et al. 2010; Warren and Mansbridge 2013.

\textsuperscript{55} O’Flynn 2006, 97; Erman 2009.
participants can count on each other to actually act on reasons they claim to share. And in anarchical circumstances, the problem of trust is particularly serious. What makes the pressures of collective action particularly severe in anarchy is precisely that, in the absence of strong institutions and thick normative consensus, opponents have reasons to question whether others will abide by their deliberative commitments. Building trust is necessary for deliberation, and it poses a challenge distinct from that posed by pluralism and disagreement.

One way to conceptualize the problem of trust within deliberative theory is think of collective will-formation as having two dimensions. Collective will-formation is what happens when participants succeed in influencing each other in deliberation oriented toward action. Deliberative influence means influence through the force of reasons—that is, through the way in which participants find each other’s reasons binding or compelling. To find a practical reason compelling in this way (that is, to endorse it as right) is to will the action that the reason prescribes, to entitle others to trust in the inferences that follow from that reason, and to hold oneself responsible for doing the same. Thus, by addressing reasons to each other, participants in deliberation can form a collective will: shared judgments about what ought to be done, and shared commitments to acting upon those judgments.56

Collective will-formation has two dimensions. Deliberative theorists talk most often about the first dimension, the dimension of reasons. From this perspective, collective will-formation is about the content of public reasons, about what, in fact, is a good reason. Indeed, reasons are what deliberation is ostensibly about. In deliberation, participants assert, challenge, justify, accept, and reject reasons. If, through this process, they come to agree that a particular

rule (or goal, or principle) constitutes a compelling reason for action, we can then speak of joint action, cooperation, or collective will-formation.

But underlying these reasons is a second, relational dimension of collective will-formation, and this is where trust comes in. Here the question is not about the justifiability of this or that reason, but rather about the more fundamental question of whether, and to what extent, the people deliberating see themselves as being accountable to each other at all. The public reasons knocked about in deliberation are only valid if there is, in fact, a public: a group of actors who relate to each other on a basis of mutual accountability, actors who recognize and uphold obligations to each other. Reasons are creatures of relationships of mutual accountability. Those relationships form the “space of reasons” within which particular reasons have meaning. Thus, collective will-formation, and the deliberation by which it comes about, are always at the same time about both the particular reasons participants take to be good and the relationships of mutual accountability that underpin the practice of offering, criticizing, accepting, and abiding by shared reasons in the first place.

Trust is an issue in collective will-formation because relationships of mutual accountability cannot be taken for granted. In political conflicts, actions based on shared reasons—and thus on the assumption that mutual commitments are being upheld—are always overshadowed by the possibility that opponents might relate to each other in purely instrumental ways. Talcott Parsons, glossing Hobbes, put the problem clearly: it is not only that people have

57 Habermas (1998, 63–64) writes of speech as having a “double structure”: in addition to its propositional content, it also affects the intersubjective relationships between speaker and listener. See also Mansbridge 2006, 108.

58 Brandom 1994, 5.

59 See also Arendt 1998, chap. 5; Arendt 2005, 189–91.

60 Erman 2010.
conflicting goals and values; it is also that “the actions of men [are] potential means to each other’s ends.” In the worst conflicts, the relationship between the parties is fully instrumentalized (or nearly so). The conflicting parties do not recognize obligations to each other, but only to themselves and their own constituencies. Collective will-formation requires the reversal of these instrumentalized relationships through the incremental recognition of mutual obligations. Understanding how deliberation relates to this dimension of collective will-formation is an important task.

### 3.2.2 Deliberation in anarchy

The second challenge deliberative democratic theory has to meet is to take account of the difference relatively anarchical circumstances make to deliberation. One of the effects of anarchy is to magnify the importance of the second, relational dimension of collective will-formation outlined above—the dimension, in other words, of trust-building. This is a problem that deliberative democrats have not yet responded to as well as they might. The reason they have not is quite understandable: deliberative democratic theory was developed primarily as a theory of the liberal-democratic, constitutional state, a theory, in other words, that assumes the presence of strong institutions of governance. But as deliberative theory attempts to move beyond the context of the liberal-democratic state, it will have to more directly confront the difference anarchy makes to deliberation.

---


62 Fully instrumental interactions are a limit case. Even in war, combatants usually recognize some obligations to each other. Walzer 1992.
In order to grasp the challenge posed by relatively anarchical circumstances, we must first recognize the work that institutions do in non-anarchical, that is, institutionally thick, settings. One of the key accomplishments of such states is to minimize the problem of trust. As I pointed out in Chapter 2, strong institutions stabilize cooperation by increasing the likelihood that rules and decision that secure agreement will, in fact, be upheld. Institutions do this through both normative and coercive means. Normatively, legitimate institutions carry with them a stock of institutional authority. Where citizens buy into the legitimacy of institutions, they will often be willing to abide by decisions even if they do not agree with them. To the extent that citizens see the process of decision-making as legitimate (as opposed to judging legitimacy purely on the basis of outcomes), the normative standing of an institution can add a normative motivational impulse toward compliance even for those who don’t agree with a decision on substantive grounds.

At the same time, strong institutions also stabilize cooperation through coercive enforcement. Law, as Habermas notes, is “Janus-faced.” Not only does it make a normative claim, by virtue of its legitimacy, on the citizen’s voluntary compliance, it also confronts its subjects as a demand backed up by force. The second, coercive face of law allows it to guarantee broad compliance. Even those who do not recognize the law as normatively binding will still comply in order to avoid sanctions.

The effect of these stabilizing forces is that in institutionally thick contexts, the second, relational dimension of collective will-formation recedes into the background. Within the context of a strong and legitimate state, people do not have to actively question and establish each

63 Tyler 2006.

other’s commitment to shared rules. For the most part, people deliberating can expect each other to play by the rules and abide by decisions that are made, and the threat of opportunistic defection is relatively well contained. In this way, institutions help unburden deliberation of the need to actively build trust. That allows deliberating citizens (and the scholars who theorize them) to focus their attention on the content of reasons, on the assumption that their basic commitments to abiding by shared reasons and collective decisions are secure.

But in relatively anarchical contexts, and especially in violent conflicts, all of that changes. Here, political opponents who would deliberate with each other must overcome the challenge of collective-will formation without the stabilizing resources, normative and coercive, available in more thickly institutionalized contexts of interaction.

The result is that in anarchical circumstances, the relational dimension of collective will-formation—the problem of trust—leaps back into the foreground. Political opponents in anarchy must directly confront the question of whether they can count on each other to observe mutual obligations or, to the contrary, are engaged in a Hobbesian free-for-all. For negotiators in relatively anarchical contexts, this is a live question in a way that (thankfully) it is not for citizens in liberal-democratic states with constitutionally established institutions that regulate and transform conflict into democratic channels. Given the vulnerability that trust entails in relatively anarchical contexts, the default judgment tends to be distrust, not trust.

66 Price 2008
67 Mitzen 2005; Mitzen 2013, 38.
68 Jervis 1978.
The bleakness of the situation, however, should not be overstated. As I argued in Chapter 2, even in relatively anarchical contexts—civil conflicts, weak states, and international politics—some standards of normative legitimacy and some rule-enforcement capability exist; hence the accuracy of speaking of relatively anarchical contexts. Furthermore, such circumstances often contain incentive structures that help to limit the risks involved in deliberating with an opponent.

For one thing, the costliness of the conflict itself can give opponents incentives to negotiate and keep agreements. “Hurting stalemates”—in which neither side can get what it wants without further escalation, but the costs of escalation would outweigh the benefits of winning—create incentives to search for mutually acceptable agreements.69 And where both sides retain the coercive capabilities to reignite costly conflict if commitments are broken, stalemates also generate incentives to abide by those agreements.70 As Robert Axelrod,71 Robert Keohane72 and others have argued, the long-run benefits of cooperation in repeated interactions, and the threat of withholding cooperation from opportunistic defectors, can lessen the temptation to defect and lower the riskiness of cooperation and truthful communication.73 Reputation effects and informal regimes can create similar outcomes, even in relatively anarchical situations.74 Finally, in some cases, costly signals can convey reliable information about the intentions of others, which helps to create the confidence needed to deliberate.75 Incentive structures like these

70 Conversely, some of the most challenging situations for building trust are those in which one party must ultimately give up weapons or other forms of coercive power (Walter 2009).
71 Axelrod 1984.
72 Keohane 2005, chap. 5.
73 But see Fearon 1995, 402.
can thus go some distance toward lowering the risk of opportunistic defection, and so make deliberation safer.

Nonetheless, despite these mitigating factors, the broader point stands: anarchy remains a tough case for deliberation, and specifically for the trust-building aspect of collective will-formation. Given the weakness of institutions in anarchical contexts, deliberating sincerely with an opponent involves overcoming a measure of risk greater than that involved in deliberating within the protections of a strong, liberal-democratic state.

3.2.3 Foreshowing a response: rebuilding the ship at sea

Where does this leave deliberation in these difficult circumstances? Deliberative democratic theory has the resources to respond to the challenge posed by anarchical circumstances. Elaborating that response will be the main work of my dissertation, and I will begin that task in earnest in Chapter 5. For now, though, I will briefly state what I see as the main potential of deliberative democratic theory to explain the emergence of rule-governed interactions in anarchy.

The way forward, I suggest, is to develop a thicker account of the trust-building capacities of deliberation. The concept of deliberation offers a way of understanding how political opponents build relationships of mutual accountability. Deliberation naturally puts people in such relationships. In offering, accepting, and rejecting reasons, people necessarily attribute responsibilities to themselves and others. As speakers, they are responsible for making valid claims: statements that are true or right. If speakers were not responsible to listeners for

---

making valid statements, those statements would quickly lose their meaning; listeners would have no reason to take them seriously. Likewise, listeners are responsible to speakers for accepting valid claims. If they were not—if, for example, the listener could reject the claim on a whim, or merely out of antipathy toward the speaker—then the speaker would have no reason to address a claim to the listener in the first place. The upshot is that when speakers address reasons to each other, they implicitly place themselves in relationships of mutual accountability: they treat each other as trustworthy users of language, and affirm their own trustworthiness. In trying to justify themselves to each other, political opponents leave behind them “a trail of commitments, obligations, and responsibilities.”

What good can shared normative accountability do? The answer is that in situations where institutionalized accountabilities are thin or non-existent—that is, in anarchy—the structure of normative accountability that deliberation generates can provide an initial foothold for rule-based cooperation to get a start. Over the long run, opponents might hope to back up those normative commitments with institutionalized incentives and enforcement capabilities. If they succeed in building good institutions, the personal bonds and commitments that emerge out of deliberation will be relieved of some of the burden of stabilizing cooperation and underwriting trust. But before such institutions are established, they need some way of breaking out of vicious circles of distrust and defection. Deliberation, I will argue, can provide that impetus toward mutual self-restraint and rule-based cooperation. To wary and distrustful opponents, the bond generated by talking to an opponent may not seem very strong, but in some cases it will be

76 Habermas 1987, 73.
77 Warren 2012a, 5.
enough to take the crucial first steps of trust-building, and thus to move, however haltingly, toward cooperation.

As I suggested in Chapter 2, the paradox of rebuilding the ship at sea is that of undertaking an activity before the seeming prerequisites for that activity are in place: the ship has to float even while it is being built. To succeed in establishing rule-based cooperation in anarchical situations, political opponents must find ways to stabilize cooperation without relying on prior institutions; the must, in effect, create new sources of stability as they go along. I will argue that deliberation, and the relationships of mutual accountability it creates, might explain this paradoxical generativity.

### 3.3 The problem of coercion

Before concluding this chapter, I want to address an important objection to the project of applying the concept of deliberation to negotiation in violent, or potentially violent, conflicts.

How does it make sense to talk of deliberation given the ubiquity of coercion—of threats, deception, and the use of force—in political conflict, especially in relatively anarchical contexts? On the one hand, to speak of deliberation while ignoring the reality of coercion would be willfully naive or even irresponsible. But on the other hand, to recognize that reality and yet still speak of deliberation is to risk watering down the normative appeal and conceptual coherence of the idea of deliberation. The overwhelming presence of coercion in political conflict, especially under conditions of anarchy, indeed poses a tricky conceptual challenge to deliberative theory. Nonetheless, I think there are good answers to these problems.

---

3.3.1 Deliberative moments in conflict resolution

First, it is important to be clear about how I mean to apply the notion of deliberation in violent, or potentially violent, conflicts. It is not that conflicting parties will interact exclusively through deliberation. Rather, it is a matter of identifying deliberative moments and deliberative effects in negotiations that also involve coercion.79 We can do so without either eroding the distinctiveness of deliberation as a mode of influence or whitewashing the real and inevitable place of coercion in conflict resolution.

Deliberative moments occur whenever conflicting parties offer reasons to each other in a way that relies upon the validity of the reason to have an effect. Such moments are not, I think, rare in negotiations. Negotiators often offer justifications of their positions, demand justifications of others, and accept or reject positions on the basis of whether they are right, fair, efficient, or otherwise appropriate.80

Admittedly, recognizing a deliberative moment for what it is can be tricky.81 Various forms of coercion, more or less subtle, often take on the guise of reasons. Sometimes “reasons” are really threats, as when a negotiator informs an opponent that if a concession is not made, he might not be able to keep extremists in his own camp from resorting to violence. Likewise, deceit and misrepresentation are usually presented in the outward form of reasons. Indeed, deceitful utterances rely upon being mistaken for truthful, valid claims in order to have their effects. As coercion becomes more subtle—especially as it takes the form of framing or

79 Warren and Mansbridge 2013, 94; Risse 2000, 18.
81 Deitelhoff and Müller 2005.
psychological cuing that subverts or bypasses conscious consideration without appearing to do so—it becomes harder to tell apart from genuine deliberation.\(^8\)

But difficult as it may be to distinguish deliberative moments from seeming deliberation, the problem is not insurmountable. In fact, the problem is resolved, or at least managed, in practice every day. Assessing each other's intentions and responding appropriately—which usually means responding differently—to sincerely offered reasons, disguised threats, and misrepresentations is something that skilled negotiators are good at. (Arguably, intuiting each other’s intentions is something that any well-socialized human being is reasonably good at.) That is not to deny the challenges of identifying and operationalizing deliberation in empirical research.\(^8\) But the fact that deception and coercion can be mistaken for deliberation doesn’t mean that reliable means of distinguishing them can’t be discovered.

3.3.2 Coercion that supports deliberation

To recognize that moments of deliberation can co-exist with coercion in negotiations helps to resolve the conceptual problem posed by the fact that coercion is widespread, and yet deliberation is antithetical to coercion. There remains, however, a separate question about legitimacy. While a mix of coercive and deliberative influence is conceivable, can it be legitimate? Scholarly interest in deliberation is driven, to a large extent, by normative values. The most prominent of these is democracy: the study of deliberation in political theory is largely

---

\(^8\) The question of sincerity poses an especially tricky problem. Is it deliberation if speakers offer valid reasons without deliberative intent—that is, if they offer the reasons only because they happen to be in their interests, and not because they are actually open to persuasion or questioning? This is a topic of ongoing discussion in both deliberative democratic theory (e.g. Warren 2007; Bächtiger et al. 2010; Calvert 2013) and IR theory (e.g. Schimmelfennig 2001; Müller 2004; Holzinger 2004; Krebs and Jackson 2007; Price 2008).

\(^8\) Bächtiger et al. 2010; Deitelhoff and Müller 2005.
the study of deliberative democracy. But as I suggested in Chapter 1, interest in deliberation is also connected to the closely related values of equality, autonomy, and mutual respect. Thus, it would be a big problem if the presence of coercion alongside deliberation marked the effective betrayal of the very values that underpin our interest in deliberation in the first place. Perhaps moments of deliberation do take place in partly coercive interactions, but when they do, one might argue that the coercion around them undermines their normative value.

That is a conclusion that I think we should resist. The presence of coercion in negotiations does not necessarily undermine the legitimacy of outcomes shaped by deliberative moments within those negotiations. Coercion can, of course, displace deliberation or render its results moot. That is a danger that deliberative democrats—no less than their “realist” critics—are well aware of. But under the right conditions, coercion can also support deliberation and in that way contribute to outcomes that are relatively legitimate.

There are a number of ways in which coercive power can support or even enable deliberation. The first is by helping to make deliberation safe. One of the difficulties of politics in relatively anarchical circumstances is that coercive power is very difficult to keep in check because it is not monopolized by a strong state. As I argued above, anarchy is a tough case for deliberation in part because of the lack of institutional safeguards and enforcement capabilities to protect rule-followers against the opportunistic defection of others. In such circumstances, the conflicting parties’ own coercive capabilities—their ability to defend themselves from each other and inflict punishments for others’ defection—can stand in for the protections normally

84 Rawls 2005; Gutmann and Thompson 1996.
85 E.g. Shapiro 1999.
provided, in functioning democracies, by the state. In short, parties who are able to defend themselves from predation or exploitation have less to fear from deliberation.

Second, as I suggested in section 3.2.2 above, the existence of a balance of coercive powers between conflicting parties can create incentives for them to engage with each other in relatively deliberative ways.\textsuperscript{86} Conflicting parties often come to the negotiating table only when they cannot get what they want by force. Indeed, for mediators or interveners in conflicts, deliberately creating a balance of power by increasing the coercive capacity of a weaker party can help to foster negotiations.\textsuperscript{87} Here again, the coercive power of individual actors in anarchical circumstances can play a role similar to that played by strong institutions in liberal-democratic states. One thing such states do is to distribute rights and empowerments—e.g. rights to vote, freedom of speech and association, transparency laws—widely enough so that decision-makers will be pushed, through citizens’ abilities to resist, to publically justify their decisions.\textsuperscript{88}

Third, coercion can support broadly deliberative forms of collective will-formation by enforcing collective decisions and thereby making them binding. As I pointed out in section 3.2.2 above, the enforcement of binding decisions is an important function of democratic institutions. It is part of the “Janus-faced” character of law, which is at the same time normatively binding and factually constraining.\textsuperscript{89} Anarchical contexts are defined by a relative lack of such institutional sanctions. But even here, the coercive power of individual actors can

\textsuperscript{86} Warren 2007, 278.

\textsuperscript{87} Zartman 2007, 19. Paradoxically, more-or-less deliberative negotiations can be thwarted by a lack of coercion. This occurs when efforts to reduce violence also make the conflict less costly to the parties involved, and thus lower their incentives to negotiate a settlement. This is one of the dilemmas of conflict management. Zartman 2007, 18; Babbitt and Hampson 2011, 52–53.

\textsuperscript{88} Warren 2007.

\textsuperscript{89} Habermas 1996, 25–29, 129–130.
play a similar role. Where political opponents apply coercive power in support of decisions they have made collectively through a relatively deliberative process, that coercion is part of enacting or instantiating the deliberatively formed collective will.\footnote{Mansbridge 2012; Mansbridge 1994.}

\subsection*{3.3.3 Coercion in a deliberative system}

The argument about deliberation and coercion that I have sketched out here is consistent with the recent turn toward “systems” thinking in deliberative democratic theory.\footnote{Parkinson and Mansbridge 2012; Warren 2012b.} A systems approach allows democratic theorists to look at the way different processes interact to create effects at the level of the political system as a whole. The whole, in other words, might be more (or less) than the sum of its parts. Since deliberative democrats are interested in democracy, they look for democratic functions or demands to be fulfilled. These demands include such things as empowered inclusion (that all who are affected by decisions have a say, and actionable rights or power to demand a say), collective will-formation (that individual preferences be brought together in such as way as to form judgments and intentions that are meaningfully shared), and collective decision-making (the ability of a collective to impose binding decisions on itself).\footnote{This three-part explication of democratic functions is from Warren 2012b. Cf. Dryzek’s (2009) three-part formulation of democratic functions based on the idea that deliberation must be authentic, inclusive, and consequential.}

But in contrast to earlier formulations of deliberative democratic theory, systems theorists do not expect these functions to all be carried out within a single deliberative forum or process, no matter how well designed it might be. Rather, they expect these functions to be best performed

\footnotesize
\begin{itemize}
\item Mansbridge 2012; Mansbridge 1994.
\item Parkinson and Mansbridge 2012; Warren 2012b.
\item This three-part explication of democratic functions is from Warren 2012b. Cf. Dryzek’s (2009) three-part formulation of democratic functions based on the idea that deliberation must be authentic, inclusive, and consequential.
\end{itemize}
by a number of different political processes working in combination, processes that include, but are not limited to, deliberation.

Deliberative democratic theory, then, is not committed to the idea that deliberation is the answer to every problem and the only legitimate political process.\(^{93}\) Indeed, there are some democratic functions that deliberation is not particularly well suited to solving. Empowered inclusion is supplied not so much by deliberation itself, as by the distribution of rights (including the right to vote) that give citizens the power to demand a role in decision-making. Likewise, decision-making itself is not always best accomplished through deliberation. Deliberation is a time-consuming and often inefficient way of making decisions, and for this reason, a democratic system will be best served by allowing some decisions to be made by other means, whether through voting, bureaucracy, or markets.\(^{94}\)

Such a systems perspective remains useful when we turn from theorizing the liberal-democratic state toward the emergence of shared rules in the anarchical contexts of international politics, civil conflicts, and weak states. A systems perspective can help to make sense of how non-deliberative forms of influence might complement, and not only displace, deliberation.\(^{95}\) I identified some of these potential complementarities above: having a rough balance of coercive power between the parties to a conflict can help to incentivize more-or-less deliberative negotiations, to lower the riskiness of deliberating with an opponent, and to enforce (using the threat of sanctions) decisions and rules that parties arrive at through more-or-less deliberative negotiations. The existence of such potential complementarities suggests that the power of

---

\(^{93}\) Warren 2012b.

\(^{94}\) Warren 2012b.

\(^{95}\) On relations of complementarity and displacement, see Mansbridge et al. 2012, 3.
deliberation to generate normatively legitimate outcomes is not necessarily undermined, and can even be supported, by the inevitable place of coercive power in negotiations in anarchy.

3.3.4 Deliberation, but is it democratic?

Finally, the talk of democratic systems and democratic functions raises the question of how democratic we should hope or expect systems of rules that emerge in anarchical contexts to be. Much of the theoretical literature on deliberation is concerned primarily with the contribution of deliberation to democracy. Democracy remains, as I suggested above, the dominant normative framework informing the study of deliberation in political science. But do these democratic aspirations and standards make sense for the study of deliberation in the tough cases of international politics, civil conflicts, and weak states?

The democratic systems approach outlined above helps, I think, to relate the deliberation that occurs in anarchical contexts to democratic aspirations. There are two points to make. First, it is quite possible to have deliberation without democracy, and in relatively anarchical contexts, that is probably what we should expect. In particular, for a system of rules or understandings to count as democratic, it needs to be inclusive, and inclusiveness is something that deliberation itself cannot supply. Deliberation is a mode of influence in which participants rely upon reasons and persuasion to coordinate their actions—but who is included within the process is another matter. Inclusion is supplied not so much by deliberation, as by distributions of power that give people who might otherwise be excluded the ability to demand that their

96 Knight and Johnson 2011; He and Warren 2011; Chambers 2012, 60.
voices be heard.⁹⁷ And in anarchical circumstances, power is often distributed in highly unequal ways. Whereas democratic states offer, at a minimum, the equal distribution of voting power, in international politics, civil conflicts, or weak states, even this minimal form of equalization is absent.⁹⁸ Particularly in violent conflicts, where the power to shape shared rules and agreements flows, in large measure, from the ability to threaten and deploy violence, there is little reason to expect that deliberation will be very inclusive.

As a matter of analytical clarity, then, it is important to keep deliberation and democracy separate. Outcomes shaped by deliberation will not necessarily be democratic ones. Nevertheless, there are ways in which deliberation can contribute to putting prerequisites for democracy in place. In anarchical circumstances, as in liberal-democratic states, deliberation functions as a means of collective will-formation, and that is a necessary, though not sufficient, condition of democracy. If democratic forms of government are to emerge out of conflicts, the emergence of shared rules and joint commitments in situations where such rules are thin or non-existent is a necessary step toward that goal.

### 3.4 Conclusion

Political opponents in anarchical circumstances who wish to negotiate cooperative agreements and shared rules face the difficult task of “rebuilding the ship at sea”—of generating new

---

⁹⁷ As Knight and Johnson (2011, 121–127) argue, deliberation (or “political argument”) is only democratic if it takes place under conditions of substantive and procedural freedom and equality.

⁹⁸ Where voting does take place in international organizations, votes are often not weighted equally, as in the vote-weighting formulas of the IMF and World Bank, or the veto wielded by the “permanent five” of the UN Security Council.
commitments, and the trust needed to act on those commitments, without the benefit of strong institutions for collective action. In this chapter, I have outlined how deliberative theory might be applied to this problem. What I described in Chapter 2 as “the emergence of shared rules and joint commitments” can be understood, in the terms of deliberative democratic theory, as collective will-formation: the process by which political actors come to establish shared judgment and intentions through the give and take of reasons.

In order to apply deliberative theory, however, we have to overcome the difficulty of cooperating outside the protections of the liberal-democratic state. We can do so, I suggest, by indentifying the trust-building potentials of deliberation—the ways in which deliberation helps to establish not only shared reasons, but also stable relationships of mutual commitment that allow people to act on those reason. I will begin that task in Chapter 5. But before doing so, I will show, in the next chapter, why we need such a deliberative theory.
When it comes to answering the question of how rule-based cooperation emerges in anarchical circumstances, we are, in a way, embarrassed by our theoretical riches. “Why do we cooperate?” is one of the basic questions in the social sciences: a question that most social or political theories—at least “grand theories”—have to answer in one way or another. As a result, there are a wide range of theoretical approaches to draw on.

In this chapter, I will survey some these approaches. My purpose in doing so is, in part, to demonstrate the need for a deliberative conception of cooperation in anarchy by showing where other approaches to understanding the topic fall short. If the goal of Chapter 3 was to argue that the concept of deliberation can coherently, and even plausibly, be applied in the difficult circumstances of political conflicts in anarchy, the goal of this chapter is to show that we need a theory like this to explain the emergence of cooperation in anarchy. In addition, by surveying the field of theoretical possibilities and placing the concept of deliberation in relation to those theories, I hope to clarify what a deliberative approach might have to offer.

To structure this survey, I will use a classic distinction between rational choice and “sociological” approaches to understanding cooperation and social order.¹ Rational choice approaches conceive of cooperation and rules as the result of individually rational, goal-directed action. Political opponents choose to make and follow rules where the benefits of doing so outweigh the costs. The problem of trust—that is, the question of whether an opponent can be

¹ Barry 1970; Elster 1989a. “Sociological” is a somewhat problematic label, as it denotes a field of study rather than any particular theory or family of theories, but it still makes sense as a category for reasons I will explain below.
relied on to act on commitments she has made— is understood as calculated risk-taking. Where they cannot be sure of what others will do, actors maximize their expected utility, using whatever information they have about the likely actions of others. Moreover, they assess whether others can be trusted to follow rules based on others’ incentives—that is, upon the assumption that others are maximizing their expected utility as well.

Rational choice models, however, are better at explaining some aspects of cooperation than others. In particular, they are very good at explaining the role and effects of power and strategic reasoning in cooperation. Bargaining models capture the way political actors shape rules by making credible threats to scuttle negotiations, and models of repeated games capture the use of monitoring and sanctions to enforce cooperative rules. But rational choice models are not as good at analyzing means of organizing cooperation other than power and self-interest, including cooperation that is based, at least in part, on intrinsic motivations and the voluntary acceptance of rules and commitments. Here, in matters of normative commitment and persuasion, rational choice models show their limitations.

*Sociological approaches* are, at first glance, better suited to explaining norm- and commitment-based cooperation. These approaches focus on the role of habit, shared meaning, and social identity in fostering cooperation and trust. They recognize that people act according

---


3 Fearon 1995; Powell 2002. I will focus, in what follows, on non-cooperative game theory and bargaining models as these are by far the most prominent in political science. For the distinction between cooperative and non-cooperative game theory, and an overview of cooperative bargaining models, see Osborne and Rubinstein 1994, chap. 13–15.


5 See, e.g., Parsons 1968. In the context of international relations, see, e.g., Katzenstein 1996; Adler and Barnett 1998; Wendt 1999.
to the “logic of appropriateness” as well as the “logic of expected consequences.” Standards of appropriateness, embedded in socially constructed norms and identities, can do much to mitigate collective action problems and foster cooperation. And when it comes to trust, social bonds of emotion and identity play a large role in shaping our assessments of what others will likely do.

But while sociological theories capture important aspects of cooperation, I will argue that they are better at explaining trust and cooperation within groups than between groups, and better at explaining the long-term stability of rule-based cooperation than its initial emergence. The challenge for these theories is to say where social resources for cooperation—the friendly identities and cooperative norms—come from, particularly in situations of violent conflict where social bonds tend to exist primarily within competing groups rather than between them. Unfortunately, socially constructed norms and identities can perpetuate conflict and enmity as well as cooperation.

To successfully account for cooperation, we need to be able to say how new social bonds and normative commitments emerge between groups—and in a way that does not presuppose the existence of the very social order we are trying to explain.

This survey thus demonstrates an important gap in our understanding of how rule-based cooperation emerges in anarchical circumstances. Rational choice models tell us much about how bargaining power, enforcement, and incentives shape cooperation. But shared rules require more than just well-arranged rewards and punishments. Cooperation is also motivated and

---

6 March and Olsen 1998.
7 Wendt 1999, 317.
8 Sztompka 1999; Fukuyama 1995; Uslaner 2008.
9 Warren 1999a.
11 Mitzen 2006.
stabilized by voluntary commitments and intrinsic, normative motivations. Sociological accounts of normative order, however, struggle to explain where these normative commitments come from in socially thin, anarchical contexts, where social bonds exist primarily within groups, not between groups. Thus, neither approach can answer the most important question: how cooperation emerges in anarchy. What we need, in short, is a theory that explains how political opponents build the kind of joint commitments and shared norms that tend to be assumed in sociological explanations of cooperation and overlooked in rationalist ones. In the chapters to come, I will explain how deliberative theory might fill this gap.

4.1 Rational choice approaches to rule-based cooperation

Rational choice approaches have contributed much to understanding how cooperation emerges in anarchical circumstances. As I argued in Chapter 2, the notion of cooperation itself can be usefully clarified by thinking, in a broadly “rational choice” way, about the arrangement of conflicting and common interests among political opponents, and about the various incentives they face in interacting with each other. Most importantly, rational choice approaches place collective action problems at the centre of their analysis, and this, I think, is essential to understanding why political opponents choose to create and maintain shared rules and cooperative arrangements in anarchical circumstances. To a considerable degree—subject to some caveats I mentioned in Chapter 2—we can think of “rules as tools,”\(^\text{12}\) as purposeful responses to collective action problems. And while the study of collective action problems is not unique to modern rational choice theory—the basic form of these problems was noticed, for

\(\text{12 Ostrom 2005.}\)
example, by both Hume and Rousseau\(^\text{13}\)—rational choice theorizing is in many ways the most sophisticated body of thinking on this topic.

### 4.1.1 Bargaining

Within a rational choice framework, the problem of how cooperation emerges in anarchy can be broken into two parts.\(^\text{14}\) First, there is a *bargaining* problem, which concerns what agreements, if any, opponents can reach through a process in which each tries to secure the best deal he can. Second is an *enforcement* problem, which concerns whether opponents can successfully maintain compliance with those agreements or rules through reciprocal monitoring and sanctions.

Bargaining models provide important insight into the processes by which political opponents bring power and resources to bear in distributional negotiations. Indeed, the basic parameters of bargaining models help to explain what bargaining power *is*. The relative power of parties in negotiations is a matter of how much they want to make a deal, that is, how valuable a deal is to each party in comparison to the status quo.\(^\text{15}\) The stronger party is the one that needs cooperation less and can thus drive a harder bargain by accepting a higher risk of the deal falling through.\(^\text{16}\)

Bargaining models also provide plausible explanations of why negotiations sometimes fail and cooperative norms do not emerge, *even when there are agreements that both sides would*}

\(^{13}\) Hume 1969, pt. 3.2.5; Rousseau 1988.

\(^{14}\) Fearon 1998. See Elster (1989a, 1–16) for a similar distinction between these “two problems of social order.”

\(^{15}\) That is the most important factor. There are others, though, including the ability to generate audience costs or otherwise tie one’s own hands, and therefore be able to make credible commitments to walk away (Fearon 1994; Putnam 1988).

\(^{16}\) This is the “war-of-attrition” model of non-cooperative bargaining. Rubinstein 1982; Powell 2002; Fearon 1998; Morrow 1999.
prefer to the status quo. One reason negotiations fail is the problem of “private information” or “cheap talk.” Each party has a “reservation price”: a minimum amount that it has to win in the bargaining in order to prefer a negotiated settlement to the status quo. The problem, however, is that each party has an incentive to misrepresent this minimum as being higher than it is, in order to win as much as possible for itself. That means that neither party can take each other’s threats to walk out at face value. And as a result, parties will sometimes unwittingly drive too hard a bargain and scuttle negotiations even though a mutually acceptable deal was available.

Negotiations can also fail because of an inability of one or both parties to make a credible commitment to upholding the agreement in the future. Commitment problems occur where the parties have reason to believe that the balance of power will shift in the future—often as a result of the agreement itself. This is a major problem, for example, in negotiations in civil wars between states and rebel groups. Peace deals usually involve inducing the rebel group to give up its arms in exchange for concessions such as political or economic reform. But once the rebels give up their arms and their power to resist, the state can renege on its promises with impunity. Because the rebels know this, they may refuse make a deal, even on favourable terms. As a result, violent conflicts can continue even when there are agreements that both sides would prefer to war.

17 Fearon 1995.
18 Fearon 1995; Powell 2006.
19 Walter 2009.
4.1.2 Cooperation and enforcement

Enforcement is the second challenge facing political opponents who want to establish shared rules. When it comes to analyzing the enforcement problem, the Prisoners’ Dilemma (or, if more than two players are involved, the “social dilemma” or “common-pool resource problem”) captures the most important features of rational choice analysis. The agreement the prisoners make not to snitch on each other in a Prisoners’ Dilemma is a paradigmatic example of an agreement that is not enforceable. Each player has an incentive to defect from this cooperative scheme. To the extent that political opponents in anarchical circumstances face Prisoners’-Dilemma-like incentives, there is little chance that mutually beneficial rules and agreements will be observed.

In response to this bleak picture, however, theorists of cooperation have pointed to two important mitigating factors. First, they argue that the repeated Prisoners’ Dilemma, as opposed to the one-shot version, may be a better model of interactions between political opponents in anarchy, interactions which generally persist through time. Importantly, repeated interactions make it possible for players to punish each other for defecting by refusing to cooperate in the future. Mutual threats of punishment change the nature of the game considerably. As long as the players value future cooperation highly enough (i.e. as long as the “shadow of the future” is long enough), and as long as they have reliable enough means of monitoring each other’s

20 Ostrom 1990. See Chapter 2, section 2.2.3 above.
actions,\textsuperscript{24} there will be strategic equilibria in which players consistently cooperate rather than defect.\textsuperscript{25}

Second, theorists of cooperation have pointed out that institutions can also do much to mitigate collective action problems. Institutions can, in effect, extend the shadow of the future by raising the benefits of continued cooperation and lowering the rewards of opportunistic defection.\textsuperscript{26} Institutions do this in many ways: they facilitate cooperation and bargaining, link cooperation across different issue areas, and carry out monitoring and sanctioning. Importantly, these institutional means of mitigating collective action problems can exist even in relatively anarchical circumstances. A key argument in this line of thought is that institutions for collective action do not necessarily have to be imposed and maintained in a top-down way by a central state\textsuperscript{27} or, in international politics, by a dominant hegemon.\textsuperscript{28} Rather, such institutions or regimes can arise in spontaneous, bottom-up ways. Political actors who find themselves confronting common problems can, under the right conditions, create institutions for self-governance, institutions that allow them to manage collective action problems even in the absence of an overweening Hobbesian sovereign.

\textsuperscript{24}Mailath and Samuelson 2006.

\textsuperscript{25}This is the conclusion generally known as the “Folk Theorem” for repeated games.

\textsuperscript{26}Keohane 2005.

\textsuperscript{27}Ostrom 1990.

\textsuperscript{28}Keohane 2005.
4.1.3 Trust

Rational-choice thinking about incentives for cooperation and rule-following also offers insight into the problem of trust. On this view, to trust is, in effect, to believe that the enforcement problem has been adequately solved. We can trust people if we believe that their incentives are properly aligned—whether through repeated games, networks, or institutional sanctions—so that they have an interest in upholding their obligations or following the rules. In other words, trust is appropriate where the trusted person’s interests “encapsulate” those of the truster.\(^\text{29}\)

Of course, political opponents usually face considerable uncertainty about whether or not others will, in fact, cooperate. Uncertainty can come from a number of sources. For one, players usually have imperfect information about each other’s preferences. Some players value cooperation more highly than others, while some are more tempted than others to defect. For example, some states might be “security seekers” while others are “expansionist.”\(^\text{30}\) Uncertainty can also be the result of an unknown future. Political opponents’ abilities to successfully threaten sanctions, and thus give others strong incentives to cooperate, can be undermined by shifts in power or technological developments. In practice, political opponents, especially in anarchy, can rarely be certain that the threat of opportunistic defection is contained. As a result, cooperation and trust always involve risk.

Rational choice models, though, provide well-developed ways to model decision-making with risk. Thus, for example, in James Coleman’s seminal model,\(^\text{31}\) the decision about whether or not to trust an opponent depends on three things: the gains (\(G\)) to be had from cooperation if the

\(^{29}\) R. Hardin 1993.


\(^{31}\) Coleman 1990.
opponent upholds the agreement or rule (i.e. fulfills trust); the loss \( L \) the truster would incur if she cooperates but the opponent does not; and the probability \( p \) that the opponent will cooperate. The rational actor, in the model, will trust if \( p/(1 - p) \) is greater than \( L/G \). In other words, she will trust if it looks like a good bet: if the probability that the other will cooperate is high enough to outweigh the harm she would incur if the other does not. That probability, in turn, is determined by the incentives facing the opponent. The more confident the truster is that opportunistic defection has been successfully deterred, or that the opponent has broadly cooperative preferences, the more confident she is that his trust will be fulfilled.\(^{32}\)

Models along these lines provide considerable explanatory leverage on the question of how and when cooperative regimes will emerge. They provide a useful conceptual framework for understanding both the demand for cooperative regimes and the difficulty political opponents face in building them.

4.2 Power and persuasion: the limitations of rational choice approaches

There is, however, an important limitation to rational choice approaches to understanding cooperation in anarchy. The problem is this: rational choice models are very good at representing strategic bargaining based on resource control, but they are not so good at analyzing cooperation that is based, at least in part, on people’s voluntary acceptance of rules and commitments that strike them as being, in one sense or another, good or right. In this part of human action, in the realm of normative commitment and persuasion, rational choice analysis reveals its limitations.

\(^{32}\) See Kydd (2005) for a model of trust-building in which actors with cooperative preferences use costly signals to identify themselves to each other.
What does it mean to say that rational choice models are good at describing power but not normative commitment? Scholars who use such models do not necessarily think of themselves as modelling power so much as decisions. But if we look at the models, the aspects of decision-making that they capture most insightfully have to do with the exercise of power in situations of strategic interdependence. By “power” I mean “A’s preferences or interests causing B to do (or changing the probability that B will do) what B would not otherwise have done through the threat of sanction or the use of force.” It is particularly the “threat of sanction” that does much of the work in game theoretic models. In situations of strategic interdependence, each player’s actions have an effect on the payoffs of others. Moreover, players deliberately use those effects—or the possibility that they might create those effects—to alter each others’ choices. An equilibrium strategy embodies precisely that kind of interactive or “best reply” reasoning: “if I do this, then she’ll do that,” and so on. Strategies work, in other words, by unilaterally changing the outcomes of the other players so as to get them to act in ways that they would not otherwise act—making a bargaining concession they would not otherwise make, or refraining

33 Johnson 2010.

34 This is how Mansbridge et al. (2010, 80) (building on Nagel [1975], Dahl, and Weber) define coercive power. It is not clear whether the power represented in game theoretical models is necessarily coercive, in part because it is sometimes deployed within the context of interactions, like market transactions, that participants can opt out of if they want to. For example, in bargaining over the price of a house, both parties try to use their bargaining power, but it’s not clear that either is coercing the other. But in situations where one or both parties can’t easily opt out of interacting with the other, the exercise of bargaining power looks more like coercion. More on this below.

35 Aumann and Drèze 2004; Morrow 1994, 79–80. This is the most commonsense way of interpreting strategic equilibria—that is, as the result of conscious, chess-like calculation. But there are significant difficulties with this interpretation, including the unrealistic cognitive and informational demands it makes on actors (Aumann and Brandenburger 1995). Many scholars thus adopt “steady state” rather than “deductive” interpretations (Osborne and Rubinstein 1994, 5) in which players do not actually engage in “best reply” reasoning. One variation of this is to interpret strategic equilibria as social norms (Landa 2006; Axelrod 1986.)
from “opportunistic” defection they would otherwise engage in. That fits the definition of power above.

The centrality of power to game-theoretic models is clear when we look at what the “moving parts” are—what really determines the outcomes in them. In the bargaining models I summarized above, what determines the distribution of rewards is the ability of the parties to make a credible threat to walk out of negotiations. Parties who have advantages in information, in their ability to generate audience costs and therefore make credible threats, or in the attractiveness of their outside options (which generally goes with having greater resources) are able to extract more concessions from their opponents. When it comes to maintaining cooperative schemes, the ability to shape each other’s behaviour by threatening to impose costs is again central. In the repeated Prisoners’ Dilemma, for example, what makes cooperation possible is precisely that each player can use the threat of withholding cooperation in future rounds to keep opponents from attempting to “sucker” them in the present round. (Without this sanction, the temptation to sucker would be overwhelming, as it is in the one-shot Prisoners’ Dilemma.)

To be sure, there are some situations in which power is not an important part of rational choice models. One is when there is a perfect harmony of interests, as in a pure coordination problem. In such situations, everybody gets exactly what he wants, and no one has incentives to defect or try to get more at the expense of anyone else. But as I argued in Chapter 2, we should

________________________

36 “Unilaterally” because in non-cooperative game theory each individual is assumed to choose purely on the basis of her own preferences, and is not obliged to work together with or uphold commitments to the others.

37 It is possible to be mislead by the fact that common cooperation in a Prisoners’ Dilemma is mutually beneficial into thinking that there is therefore no need for players to threaten punishment. But as the Folk Theorem makes clear, cooperation is only possible if each player can threaten to hold the other to a lower payoff (her “minmax” value) than what she would get by suckering the other in the present round. Morrow 1994, chap. 9.
generally expect political conflicts to involve conflicting interests as well as common ones. As long as we are talking about situations in which there is a significant distributional dimension to bargaining, power will play a central role.

Another situation in which players in a game do not exercise power over each other is in competitive markets. Here, the parties do have conflicting interests: buyers want a low price and sellers want a high price. But buyers and sellers have no power over each other because if either threatens to walk out or withdraw cooperation, the other party can easily find someone else willing to buy or sell at the market price. Political conflicts, however, are generally not like competitive markets. Israelis and Palestinians, for example, are stuck with each other; neither can easily walk away and find someone else to “do business” with.

The fact that most political opponents cannot easily opt out of interaction with each other also fits with the common use, in many game-theoretical models, of negative inducements over positive ones, i.e. of punishments rather than rewards. Many strategies can be interpreted in either positive or negative terms. Is a person playing a “grim trigger” strategy in a repeated Prisoners’ Dilemma “threatening” to defect if the other defects or “promising” to cooperate as long as the other cooperates? Within a market context, we tend to emphasize the rewards—after all, if the exchange were not rewarding, the parties could just do business somewhere else. But in situations of inextricable strategic interdependence, the negative interpretation seems more appropriate—as the standard name for this strategy, “grim trigger,” reflects. Thus, when

__________________________

38 Knight 1992; Knight and Johnson 2011.

39 In a “grim trigger” strategy, the player cooperates in the first round and continues to cooperate as long as her opponent cooperated in the previous round. But if the opponent defects, she then defects for the rest of the game. Morrow 1994, 266.

40 The “negative” terminology of punishment, threats, monitoring, and sanctions is characteristic of the literature on repeated games. E.g. Morrow 1994, chap. 9; Osborne and Rubinstein 1994, chap. 8; Mailath and Samuelson 2006.
it comes to analyzing most political conflicts— which, as I argued in Chapter 2, involve both common and conflicting interests, as well as incentives to defect from cooperative schemes— rational choice models are well-suited to capturing power dynamics, and especially power that comes from the ability to impose adverse outcomes on others.

What rational choice models are not very good at representing, however, is the role of normative commitments and persuasion in establishing and maintaining rule-based cooperation. Real-life bargaining and negotiation rarely takes the form of “pure” strategic bargaining, in which opponents interact only through threats and promises.\textsuperscript{41} Normative commitments also play a significant role in cooperation, even in relatively anarchical circumstances.\textsuperscript{42} That goes for both the processes through which new rules or cooperative regimes are established and for their ongoing maintenance. The negotiation of new regimes doubtless involves power-based bargaining, but it is also shaped by prior normative standards and precedents, which constrain the terms within which bargaining takes place.\textsuperscript{43} Likewise, normative commitments also play a role in maintaining compliance. Of course, some of the motivating force of social norms can itself come from sanctions. As I argued in Chapters 2 and 3, institutions that back up normative commitments with force play an important role in stabilizing cooperation. Nevertheless, it is also clear that some institutions—and perhaps the most successful ones—also rely on intrinsic motivations.\textsuperscript{44}

\textsuperscript{41} Elster 1989a; Deitelhoff and Müller 2005; N. C. Crawford 2009.

\textsuperscript{42} Deutsch 1957; Katzenstein 1996; Wendt 1999.

\textsuperscript{43} Price 1998

\textsuperscript{44} Frey 1994; Tyler 2010; Mansbridge 2009b.
One might object that normative or intrinsic motivations can themselves be incorporated into rational choice models as payoffs or preferences. Behavioural game theorists, for example, have had some success in testing models of bargaining situations in which players have an “inequality aversion”—an in-built preference for equality that manifests itself in being willing to reject unfair divisions even if doing so means getting nothing. Similarly, norms of reciprocity can be modelled, in which people gain utility from the gains and losses of others, depending on whether those others have helped or hurt them (or even third parties) in the past. Generally speaking, any normative motivation can be incorporated into a rational choice model by simply adding or subtracting utility for those strategies in which the player either obeys or breaks the norm. And particularly if these normative costs and benefits are applied immediately, without the violation needing to be detected by others, they can plausibly be interpreted as representing non-coercive, intrinsic motivations, as opposed to coercively applied social sanctions.

Nonetheless, such models have struggled, I think, to gain traction in political science, in comparison to models in which actors have more straightforwardly “selfish” preferences. (Formal and experimental work on normative preferences seems to take place more often in economics departments, Elinor Ostrom and her colleagues being a notable exception.) I suspect that one reason they struggle is that when normative preferences are incorporated into rational choice models, it is not clear what explanatory work the models themselves are doing. There

45 Fehr and Schmidt 1999
46 Falk and Fischbacher 2005; Fehr and Fischbacher 2004a.
48 E.g., the work of Samuel Bowles, Ernst Fehr, Urs Fischbacher, Herbert Gintis, and Colin Camerer.
49 Evolutionary game theory (e.g. H. P. Young 1998; H. P. Young 2004; Bicchieri, Duffy, and Tolle 2004; Aumann 1997) is something of an exception. But these are better understood as “sociological” approaches because in them, individuals make choices by following norms or rules rather than strategic reasoning. See note 59 below.
remains a suspicion among political scientists who use game theory that to explain cooperation by pointing to cooperative preferences is not so much an explanation as the short-circuiting of explanation.\textsuperscript{50} Such preferences exist, of course, but then the real question is about where those preferences come from or why they show up in some places but not others. And on this question, the models themselves do not have much to say because within the terms of any given model, normative preferences, like all preferences, are simply given.

In sum, rational choice approaches are most at home, and most effective, in theorizing the role of power and extrinsic incentives in the emergence and maintenance of cooperative schemes. But their lack of insight into normative commitment and persuasion ultimately leaves them unable to explain crucial aspects of how cooperation emerges under anarchical circumstances.

4.3 “Sociological” explanations: Culture, norms, and identity

Sociological approaches are the second broad body of literature on how rules and cooperation emerge in anarchical circumstances. These are approaches that focus on the effects of shared meaning, norms, and social identities. “Sociological” is a somewhat problematic label, as it denotes a field of study rather than any particular substantive theory. (There can be, for example, rational choice social theories, such as James Coleman’s.\textsuperscript{51}) But the sense of the label is that these approaches to understanding cooperation involve thinking of actors as forming, in one way or another, a “society,” along with socially constructed meanings, norms, and identities.

\textsuperscript{50} Morrow 1994, 19. I will return to this theme in Chapter 5.

\textsuperscript{51} Coleman 1990.
What answers do these approaches offer to the question of how cooperation and trust emerge in anarchical circumstances? First, they work from a different conception of human action and motivation than do rational choice approaches. They argue that people often follow a “logic of appropriateness” rather than (or in addition to) a “logic of expected consequences.”

That is, in deciding what to do, people don’t only ask themselves what outcome they want to achieve; they also ask themselves what kind of action is right or appropriate, where “right” or “appropriate” refer to prescriptive, that is, normative standards. In some cases, what is “right” may, in fact, be to do whatever is needed to achieve a particular outcome—and in those cases, the two logics overlap. But in many cases, the standard of appropriateness people apply to their actions is not, or not solely, consequentialist, but moral, legal, or role-related instead. The answer to the question, What should I do? often comes from asking, What is the right thing to do for me as an x? where x is a socially constituted role such as a state, a representative of a people or a cause, or a leader of a party.

Importantly, action oriented toward appropriateness can do crucial work in overcoming the collective action problems frequently posed by cooperation in anarchy. In the one-shot Prisoners Dilemma, for example, actors with a purely instrumental action orientation will always defect, which means that both receive the third-best payoff. But if the actors subscribe to a norm of cooperation, and allow that norm to govern their choices, the result will be mutual cooperation, the second-best payoff for each. Notice that this is different from merely re-jigging

---

52 March and Olsen 1998.
53 Parsons 1968, 44–45, 76.
54 Something like this is expressed in the tradition of political realism. Niccolò Machiavelli 1998; Weber 1946; Jackson 2000.
55 Wendt 1999.
the payoffs so that both parties prefer mutual cooperation to suckering the other.\textsuperscript{56} It is a different mode of action, in which choices are made categorically, according to the kind of action that is appropriate, not according to outcomes.\textsuperscript{57} Normative motivations, then, can help to mitigate collective action problems. To the extent that people are committed to upholding cooperative norms, they may resist the temptation toward opportunistic defection that bedevils many cooperative schemes.

Furthermore, normative motivations can sometimes sustain cooperation even if not all the people involved subscribe to the norm. Most systems of norms include norms enjoining the punishment of rule-breakers. The threat of being punished then gives others an instrumental reason to follow the norm.\textsuperscript{58} Normative motivations, in other words, can combine with extrinsic motivations and coercion to create “motivational ecologies”: equilibria including stable populations of normatively and non-normatively motivated actors who sustain high levels of cooperation.\textsuperscript{59}

Second, sociological approaches to understanding cooperation point to the fact that the interests that political opponents pursue are themselves socially constructed. Socially constructed interests can change, and that represents a possible path to the emergence of cooperation and new systems of rules. In international relations, for example, it matters a great deal how “national

\textsuperscript{56} Cf. Bicchieri 2006, chap. 1.

\textsuperscript{57} Sen 1977; Chapman 2003; Heath 2008. I will have much more to say on this topic in Chapters 5 to 7.

\textsuperscript{58} Parsons 1968; Heath 2001a.

\textsuperscript{59} Bicchieri, Duffy, and Tolle 2004; Bendor and Swistak 2001. This research uses the tools of evolutionary game theory and computer simulations to find stable distributions of different, more and less norm-conformative, strategies. Note, however, that this kind of game theory is not really about “rational choice” at all. The actors in these models do not choose their strategies in order to maximize utility. Strategies, rather, are determined by much simpler learning mechanisms, or, in the case of biological evolutionary models, by genes and natural selection. Aumann 1997; H. P. Young 1998; H. P. Young 2004.
interests” are constructed. States that define their interests in terms of military glory, imperial expansion, or the territorial unification of an ethnic diaspora are inherently more conflict prone, since their interests conflict directly with the core interests of other states. On the other hand, states that define their interests in terms of open trade and multilateralism are much more likely to arrive at cooperative arrangements. While these interests are likely determined in large measure by domestic factors, interests can also be shaped “horizontally” through the socialization of states by other states or by transnational civil society.

Third, group identities can also help to contain conflict and bolster cooperation. Identity is often understood to be the deepest level of social construction and influence on the self. Whereas interests and norms can, to some extent, be consciously chosen, identities, presumably, cannot—or at least not in the same way, since one’s identity defines the very self that would do the choosing. To the extent that adherence to rules or cooperative schemes becomes built into the way actors see themselves, cooperation will likely become quite resilient. This, for example, has occurred within “security communities,” such as NATO, and, arguably, among all democracies. War between Canada and the United States, for example, has become nearly unimaginable. The possibility is ruled out not by rational calculations; instead, it is simply incomprehensible given the way Canada and the United States (both their citizens and leaders)

60 Moravcsik 1997; Jervis 2002.
61 Rosecrance 1986.
63 For the idea that identities can be consciously chosen for instrumental purposes see Fearon and Laitin 2000; Coleman 1990. For a different, Kantian conception of identity and self-construction, see Korsgaard 2009.
64 Deutsch 1957; Owen 1994; Risse 1996; Adler and Barnett 1998.
think about themselves and each other. Security cooperation between Canada and the United States has arguably become so engrained as to simply be habit, something taken for granted.

One of the most important effects of identity lies in demarcating group boundaries. Cooperation and trust are bolstered when people see themselves as belonging to the same “in-group,” particularly when that group is defined in contrast to an excluded “out-group.” In-group/out-group bias is a well-established finding in social psychology. It is often combined with the “attribution bias”—the tendency to interpret the actions of one’s own group (or oneself) in a positive light and the same actions, when taken by others, in a negative light. When faced with the challenge of making and following rules together, affirming a shared identity may go a long way toward stabilizing cooperation. For example, Franco-German reconciliation after the Second World War was accompanied by a deliberate effort to foster a common European identity, an identity that was given institutional expression in the development of the European Union and its precursors.

4.4 Cooperation and the emergence of social order

One of the challenges facing these sociological approaches to explaining cooperation, however, is to account for where cooperative norms and shared identities come from. Part of what makes it so hard to establish shared rules and commitments in anarchy is precisely that these normative and social resources for cooperation—the norms, friendly identities, and social

65 Wendt 1999, 289.

66 A classic experiment here is the “Robbers’ Cave.” Sherif et al. 1988.

67 More specifically, people tend to attribute their own success to personal dispositions or virtues and the success of others to situational factors like good luck. Morrell 2010, 102–4. See also Jervis 1976, chap. 3.
relationships—tend to be missing. That, as I argued in Chapter 2, is part of the challenge of “rebuilding the ship at sea.” In the most difficult situations, opponents do not see each other as friends or rivals, but as enemies.\(^{68}\) And while they may indeed follow norms, the norms they follow are often particularistic, not shared: norms that prescribe loyalties to their own group and not obligations to each other. In many violent or potentially violent conflicts, social and cultural dispositions contribute as much to conflict as to cooperation.\(^{69}\)

If sociological theories are to explain how cooperation emerges in these difficult circumstances, then, they need to provide an account of how normative orders change—and why they should change in the direction of supporting intergroup cooperation rather than enmity. If the social forces of shared meaning, norms, and identity are to foster cooperation between groups in conflict, those groups must, in effect, form a society—a layer of social order at the intergroup level, as opposed to the social order that exists within each group. Importantly, the social change or learning this involves is not simply a change in the content of social norms and shared meaning. Rather, it is also, and more fundamentally, a change on the level of people’s commitment to that social order itself.

To illustrate the nature of the problem, consider the disputes about “parity of esteem” in Northern Ireland.\(^{70}\) “Parity of esteem,” is the principle that the two communities in Northern Ireland, unionist and nationalist, must be recognized and treated as equals. It emerged as a central principle of the peace process and a guiding principle in the design and operation of the new power-sharing institutions. While the principle itself is widely accepted, its implications are

\(^{68}\) Cf. Wendt 1999, chap. 6; Mitzen 2006.

\(^{69}\) Heath 2014b, chap. 3.

\(^{70}\) O’Neill 2003.
contested, the most controversial being perhaps the renaming of the Royal Ulster Constabulary (symbolically unionist) to the Police Service of Northern Ireland (neutral). Notice, however, that such controversies can only take place if there are social norms and standards of judgment that bridge the two communities. The controversy presumes, in effect, the existence of a Northern Irish society comprising those communities, and some degree of allegiance to that society on the part of those who contest those norms.

But not everyone in Northern Ireland recognizes such a society. From the perspective of hard-line republicans, for example, there are not “two communities” in Northern Ireland; there is only one community: the Irish Republic, which has a right to sovereignty over the whole island of Ireland and the loyalty of all the people who live there. Unionists who opposed the unification of Ireland are not, on this view, a community with legitimate national aspirations. From this perspective, as from the corresponding hard-line unionist perspective, there is no “Northern Irish” society in which both groups are members on equal footing, nor do the norms of that society—including the principles of mutual respect and the parity of esteem—have any purchase. The challenge, then, for sociological accounts of cooperation is to explain how conflicting groups shift toward seeing themselves as part of a shared normative community or society.

When it comes to answering this challenge, most sociological accounts focus on identity change. Intergroup norms and cooperation begin to take hold when political opponents change the way they see themselves in relation to the “Other.” On this view, the crucial shift that takes place is that opponents come to see themselves as sharing a common identity—seeing

---

71 Mulholland 2002, 65–70.
themselves, to use the example above, as “Northern Irish,” as opposed to simply unionist or nationalist. It is this shared identity that provides, in effect, the motivational anchor for shared rules and cooperative norms. People comply with norms because compliance is part of what it means to be an x, whether x means “Northern Irish,” “NATO member,”73 or “responsible international citizen.”74 That is, where specific norms are constitutive of an identity, adopting that identity means that those norms become binding for the actor in question. Furthermore, to the extent that political opponents come to see themselves as sharing a common identity, they may shift toward more altruistic actions. As Alexander Wendt argues, when actors see themselves as part of a new “we” they will “make the welfare of the group an end in itself, which will in turn help them overcome the collective action problems that beset egoists.”75

How are such common identities constructed, and how do they change? The central mechanism of identity change is “social learning”: a process through which actors internalize roles that others ascribe to them in interaction.76 Actors learn their identities from each other. In each “scene” of interaction, the person acting (“Ego”) assumes, and tries to impose on the other (“Alter”), an understanding of their shared situation and respective roles. Alter then tries to make sense of Ego’s action, and responds out of her own understanding of the situation and their roles. Through successive iterations, Ego and Alter eventually move toward shared understandings of the roles they are playing.77

73 Risse 1996.
75 Wendt 1999, 337.
76 Wendt 1999, 327 ff.; Adler and Barnett 1998, 43–45; Aiken 2010, chap. 2. This work draws on the “symbolic interactionism” of George Herbert Mead. See also Habermas’s (1987, vol. 2, chap. 5) communicative reinterpretation of Mead’s theory.
77 Wendt 1999, 330–32.
Social learning thus provides an explanation of how shared cooperative or rule-abiding identities might emerge out of patterns of cooperative interaction and mutual exchange. In this respect, the constructivist account of how cooperation emerges bears some resemblance to the older idea (going back at least to Kant\textsuperscript{78}) that patterns of mutually beneficial exchange, such as commercial trade, discourage war. The important difference, however, is that social identity theory offers an account of how such cooperation, which is initially motivated by self-interest, can shift toward internalized norms and intrinsic motivations.\textsuperscript{79} The “transactions” and mutual exchanges that are the engine of this identity change may take many forms,\textsuperscript{80} but international organizations and institutions are often seen as especially important venues for socialization.\textsuperscript{81}

This constructivist account of how cooperative norms emerge gets some important things right. In particular, it is right to focus on the effects of interaction and social relationships on how (and whether) people orient themselves toward social norms. In that regard, it shares a point of connection with the deliberative theory I develop in the chapters to come.

But the constructivists’ more particular focus on identity as a way of theorizing these interactive and relational effects also has significant drawbacks. First, there are reasons to be wary of relying on identity to explain cooperation in contentious or “intractable” conflicts, especially in the early stages when cooperation is first emerging. Many of the most difficult conflicts are characterized by frankly oppositional and mutually exclusive identities. Identities, moreover, are generally understood to be relatively deep-set, difficult to change, and closed to

\textsuperscript{78} Kant 1991a. See also Jervis 2002.

\textsuperscript{79} Wendt 1999, 334.

\textsuperscript{80} Adler and Barnett (1998, 41) define these “transactions” widely as any “bounded communication between one actor and another.” See also the earlier formulation of Deutsch 1957.

\textsuperscript{81} Adler and Barnett 1998, 41–43; Johnston 2001; Checkel 2005; Bearce and Bondanella 2007.
compromise, in comparison to preferences and interests. If cooperation depends upon these identities first changing, the prospect for cooperation in many anarchical situations looks very dim. That may be the grim reality, but it may also be too pessimistic and restrictive. It is possible to imagine political opponents beginning to cooperate without necessarily adopting a common identity, seeing each other as friends, or (even less plausible) engaging in altruism. Indeed, it is likely, especially in early stages of conflict resolution, that challenges to particularistic or exclusionary identities may even be counterproductive.

A second problem with identity-based explanations of cooperation is that the mechanisms of social learning and role internalization require that participants are engaging in relatively cooperative ways to begin with. Social learning tends to create “self-fulfilling prophesies”; as people internalize the roles implied in the way others treat them, social learning tends to stabilize existing patterns of interaction. That suggests, however, that if cooperative or rule-abiding identities are to be learned, some amount of cooperative interaction must already be taking place before the identity-affirming mechanism can kick in.

In fact, a closer look at constructivist theories confirms this suspicion. When it comes to explaining the initial emergence of cooperation, many constructivists end up appealing to mechanisms other than identity change. One important mechanism is self-restraint. In Wendt’s account of how friendly identities emerge, self-restraint is the crucial “enabling” or “permissive” variable, and “the ultimate basis for collective identity and friendship.” It is by exercising self-

---

82 See, e.g., Elster, Ofte, and Preuss 1998, 251.
84 Wendt 1999, 331.
85 Wendt 1999, 343, 360.
restraint that opponents demonstrate to each other that they recognize each other’s rights and needs, and it is the consistent application of self-restraint (sometimes unilaterally at first) that builds trust. Similarly, Charles Kupchan’s account of how “stable peace” emerges begins with unilateral acts of self-restraint, followed by reciprocal self-restraint. These actions are driven not by shared identity, but by a self-interested desire to avoid costly conflict. Shared identities emerge only much later, after patterns of cooperation are well established. Similarly, Adler and Barnett suggest that while trust and collective identity are mutual reinforcing, “trust logically comes prior to identity. Once some measure of trust develops, however, a collective identity is likely to reinforce and increase the depth of trust.” In other words, once cooperative interactions are up and running—and in particular, once political actors have begun to establish shared “practices,” institutions, and “self-restraint subjectivities”—shared identity makes sense as a mechanism for reinforcing and stabilizing these normative commitments. But as a way of explaining the initial emergence of cooperation, the concept of identity appears to be less helpful.

A third difficulty is that the mechanism of social learning can entrench enmity just as easily as friendship. Constructivists are well aware of this fact, and indeed use it to explain how a “Hobbesian” culture of mutual suspicion and “self-help” can perpetuate conflict in international politics. When political opponents cast each other in the role of “enemy,” social learning will

86 Wendt 1999, 357–63. See also Adler and Barnett 1998, 43; Adler 2008.
87 Kupchan 2010, 35–52. See also Wendt 1999, 342.
88 Adler and Barnett 1998, 45–46. See also Rathbun 2012.
89 Adler 2008; Adler and Pouliot 2011.
90 Adler 2008, 197-98.
91 Wendt 1999, 332.
tend to reproduce and internalize enemy identities; conflict and enmity can be self-fulfilling prophesies, too.\textsuperscript{92} The problem is that the symbolic interactionist logic does not itself offer any reason to think that interactions will go one way or the other—and, indeed, if people start out interacting in mutually destructive ways, social learning will tend to entrench that pattern. That is not to say that the pattern cannot change, but it does suggest that the impetus for change will not be social learning itself.

To be clear, these difficulties do not pose an objection to identity-based explanations of cooperation or norm compliance in general. Indeed, as I argued in the previous section and in Chapters 2 and 3 above, mechanisms of socialization and the construction of a shared “lifeworld” are important means of stabilizing both cooperation and deliberation. And later in the dissertation I will point to a number of ways in which such mechanisms might complement the work that deliberation does. The argument I wish to make, rather, is that identity works better as a way of explaining how patterns of cooperation, once established, are stabilized and entrenched than how they get started in the first place. For example, shared identity is a plausible means of stabilizing peace within a security community: a group of states among whom war has become unthinkable.\textsuperscript{93} But by the time such a security community has formed, much important cooperation and normative change must already have taken place. (For example, according to Adler and Barnett, even a “nascent” security community, in the earliest stages of development, is already characterized by significant security cooperation or alliances.\textsuperscript{94})

\textsuperscript{92} Wendt 1999, 339; Mitzen 2006; Price 2008, 205–06; Checkel 2011.

\textsuperscript{93} Risse 1996.

\textsuperscript{94} Adler and Barnett 1998, 52.
Furthermore, identity offers persuasive explanations of how social norms gain a motivational grip on actors within already-established groups. In addition to the social learning (or symbolic interactionist) mechanism described above, membership in a group makes actors susceptible to the forces of imitation (or conformism) and social pressure. Actors who are anxious to become members of a group, or to solidify their status within that group, will tend to adopt those norms, either because they see them as part of being a group member (imitation) or because they fear disapprobation (social pressure). These are important mechanisms for the diffusion of norms among states, including human rights norms and the recent emergence of a norm against the use of anti-personnel landmines. But here again these identity-based mechanisms are most likely to apply in situations where actors already share, or aspire to share, a common, cooperative identity. Their very reliance on shared identity, though, makes it difficult to see them as explanations for the emergence of cooperation among people who do not see themselves as belonging to a common “in-group,” and who therefore care little for each other’s praise or blame.

Finally, I should point out that there is a sense in which group-formation does play an important role in the deliberative theory of cooperation I will develop in the chapters to come. I will argue that the central contribution deliberation makes to the emergence of cooperation is the formation of a particular kind of group: a public. A public is a group of actors who relate to each other on a basis of mutual accountability—actors, that is, who take themselves to be responsible to each other in normative terms. When people see themselves as part of a public, their own reasons for acting (in effect, their preferences) become responsive to shared standards of


justification or shared norms. Moreover, actors who are part of a public can see themselves as forming a collectivity, a “we,” for the purposes of acting and judging together. That is part of what it means to speak of collective action and collective will-formation. (I will offer a full account of this concept of the public in Chapter 6.)

The concept of the public thus bears some resemblance to the thicker, identity-based communities theorized in sociological accounts of cooperation. But there are also some important differences. Members of a public do not necessarily see themselves as sharing a common identity. Publics can form among actors who, in Habermas’s phrase, “concede one another the right to remain strangers.” That is because the social bonds that form a public are not sameness or homogeneity, but rather the recognition that each other has rights and legitimate needs and a commitment to self-restraint. Publics form not out of shared identity so much as from an awareness of sharing a common fate and facing common problems that call for collective responses.

To be sure, once a public forms, membership in that public may later come to be anchored in actors’ identities. If the formation of publics does make possible new forms of collective action and cooperation—the central claim of my dissertation—then we should expect those cooperative patterns of interaction to be reinforced and stabilized through the “sociological” mechanisms of social learning and internalization outlined above. Publics, and the

98 As I pointed out just above, it is interesting that some constructivist scholars find themselves falling back on similar terms—mutual recognition, self-restraint, common fate—when attempting to explain how cooperation emerges.
99 Mitzen 2013, 9–10; Erman 2013, 862; M. S. Williams 2009; cf. List and Koenig-Archibugi 2010, 83. See also Kelman 2005, 646.
rule-base cooperation they support, may thus be stabilized through changes in identity. But shared identity is not essential to the formation of a public—and that is a good thing because it allows deliberation and cooperation based on publicity to get a start even in socially thin environments.

The broader critical point, then, is that when applying sociological theories of cooperation to anarchical circumstances, it is important not to presuppose the existence of the very social order—the norms, friendly identities, and social relationships—the emergence of which needs to be explained. Without an account of where those solidarities and normative bonds come from, sociological approaches cannot explain the most important thing: how new cooperative orders emerge out of anarchy.

### 4.5 Conclusion

Both rational choice and sociological approaches offer valuable theoretical resources for understanding how cooperation emerges in anarchy, but they also leave us with a critical theoretical gap. On the one hand, rational choice approaches provide valuable models of how political opponents deploy power and structure incentives in both the negotiation of rules and their later enforcement. But these models are not very well-suited to explaining the consensual or voluntary aspects of cooperation. Successful schemes of cooperation depend not only on extrinsic motivations, on the fear of sanctions and hope of rewards, but also on intrinsic motivations, on the belief that following the rules is (by one standard or another) right. On the other hand, sociological theories that begin with normatively motivated action face the challenge

---

100 Cf. O’Flynn 2006, chap. 3.
of accounting for where normative resources for cooperation—cooperative norms and shared
identities—come from, particularly if the starting point is a socially thin, anarchical situation. As
a result, neither approach gives us the grip we need on the critical question of how political
opponents *build* new voluntary commitments and relationships of social accountability. In the
chapters that follow, I will argue that deliberative democratic theory offers a promising way to
fill that gap.
Chapter 5  Acting for a Reason

According to the theory I am developing in this dissertation, deliberation—the give and take of reasons—can be an effective way of bringing about collective action in anarchy. But for that to be true, reasons have to be able to get a motivational grip on the people who hear them. If arguments are to matter in politics, reasons must ultimately be able to affect actions.

Saying how reasons get a motivational grip on people, however, is not easy. The problem, I would suggest, is that our intuitions about this are divided between separate, and seemingly incompatible, conceptions of practical rationality: the normative and the instrumental; the “logic of appropriateness” and the “logic of expected consequences”1; homo sociologicus and homo economicus.2 Both understandings of practical rationality can lay claim to an intuitive plausibility. On the one hand, there does seem to be a basic normativity or “oughtness” about reasons. Good reasons, it seems, should be motivating. To be confronted by what you take to be a good reason and yet be wholly unmoved by it is in some sense not to experience it as a reason at all.3 Of course, you might choose not to act on a good reason because there are other better or more important reasons weighing against it, but that is another matter. For any particular reason taken on its own, to understand it as a reason seems inherently to acknowledge it as making a claim on you, a claim about what you should, or should not, do or believe.4

______________

1 March and Olsen 1998.
2 Elster 1989a, 97.
3 The philosophical position I am expressing here is known as “internalism.” Korsgaard 1986; T. Nagel 1970; B. Williams 1981.
4 Notice, moreover, that this applies to all reasons: not only ones that are “normative” or “moral” in content, but “instrumental” ones as well. If action x is the best way to achieve my goal y, then I should, rationally, do x (again putting aside the possibility of my having other reasons to the contrary).
But on the other hand, it seems equally clear that rationality requires choosing the right means to achieve your ends. Instrumental rationality seems to capture a compelling intuition, namely, that practical rationality is about being effective in the world. Whatever else rationality implies, it is surely true that rational actions should not be “self-defeating,” to use a phrase often deployed, to good rhetorical effect, by rational choice theorists. Moreover, many scholars are skeptical that simply acknowledging that they ought to do something is enough to motivate people. To be motivationally effective, many suspect, reasons must ultimately be connected to the interests of the actors whose reasons they are.

In this chapter, I will argue that this apparent split between normative and instrumental rationality is misleading. Practical reasons are both normatively prescriptive and motivationally effective. A reason just is a motivation that the actor endorses as right or good. Thus, if an actor recognizes herself as having a reason to act, there is no need to posit an additional motivational mechanism (such as self-interest) to explain why she acts on the reason. Reasons are inherently motivating. Furthermore, that fact is important because it means that reasons can be an answer to the incentive problems that often afflict collective action.

The argument will unfold as follows. In section 5.1, I will present a two-part or reflexive conception of what it means to act for a reason. To act for a reason means, on this view, to take the incentive or principle on which you act to be a good one, and to be motivated by your perception of its goodness. Most importantly, acting for a reason does not mean acting for a

---

5 See Rubinstein (2000, 87), who is critical of this phrase.
7 Gauthier 1991.
8 Korsgaard 2008b.
“normative” motivation instead of some other motivation; rather, it means that you take the motivation you do have to be a normatively binding one. Moreover, the normativity of practical reason does not imply a disregard for consequences. To the contrary, the instrumental imperative to be effective in the world is integral to this conception of practical reason.

In section 5.2, I argue that the apparent tension between the normativity of reasons and instrumental rationality is also due to misunderstandings about instrumental rationality. Construing rationality in terms of effectiveness alone distorts the nature of the real choices people make, which often involve choices between ends, as well as between means. The idea of maximizing utility obscures this aspect of choice, often by implicitly assuming that a person’s ends are in harmony.

Finally, in section 5.3, I apply these ideas about practical rationality to collective action problems. Collective action problems are sometimes seen as posing a battle between conflicting normative and instrumental “logics” or “rationalities,” but that is a mistake. These are not clashes between different types of rationality, but simply conflicts between reasons, some of which counsel cooperation and others defection. Rationality consists, first, in adjudicating between these conflicting motivations and deciding what is in fact the right thing to do and, second, in putting that judgment into effect.

Defending this conception of practical rationality is a crucial step toward developing a broader theory of deliberation and cooperation in anarchy. Good reasons are inherently motivating, and that, ultimately, is what makes it possible for the give and take of reasons in deliberation to affect individual choices and (as I will argue in later chapters) collective action.
5.1 A two-part conception of acting for a reason

Let me begin by restating the problem that hangs over the idea of rationality: the apparent split between normative and instrumental conceptions of rationality, both of which lay claim to intuitive plausibility. This divide is due to mistakes in the way both elements in this duality are understood: both the normativity of reasons, and the conception of instrumental rationality that is often set opposite it. I will start, in this section, with the normative side.

The heart of the seeming tension between normative and instrumental conceptions of reason is that the normativity of reasons—their “oughtness” or prescriptiveness—seems disconnected from the real sources of people’s actions and motivations. Normatively motivated action is often understood as acting for the sake of duty: acting, somehow, in a very principled, morally rigorous, Kantian way. To be normatively motivated is to do the right thing, come what may.

If that is what the normativity of practical reasons implies, we are rightly skeptical of whether people are really like this. There are at least two problems. First, how often do people really act for the sake of duty? Are they really so disinterested? While there may be the occasional true “Kantian” out there, there surely aren’t many, especially in the grubby world of politics, in which states seek power, politicians seek re-election, interest groups seek economic advantage, and voters seek government largesse. Nor would it even be good if people were motivated only by a sense of duty, as this would exclude many other motivations we value, such

9 Elster 1989b, chap. 6.
as affection and concern for others. But these objections are based on a mistaken understanding of what is implied in the normativity of practical reasons and of what it means to act for a reason. The normativity of reasons does not consist in supplying a separate “normative” motivation to be set against others. Rather, the normativity of reasons is reflective and formal: it consists simply in the actor’s reflexive endorsement of her motivation as a reason, as normatively binding or compelling.

To explain, a practical reason can be thought of as having two parts: (1) the actor’s motivation itself, whatever it is that urges the action, so to speak, upon the actor, and (2) the actor’s endorsement of that motivation as good, that is, as providing a reason to take the action it suggests. These two parts of motivation can be thought of as corresponding to passive and active aspects of human subjectivity. “Passive” suggests the sense of being acted on. We find ourselves subjected to a wide assortment of motivations—inclinations, desires, or considerations of any sort—that press themselves upon us, urging us to take one action or another. Some inclinations appear to be basic instincts: desires for food, water, shelter, sex, bodily safety, companionship. Some inclinations may be selfish, others altruistic. Importantly, although many of these motivations can be thought of as desires or objects of desire, they can also be values or principles. These are all forms of what Hume called “passions”: a term that includes not only emotions but anything that could stir us toward action, including such things as principles of

---

12 Korsgaard 2008b, chap. 6.
13 Korsgaard 2008b, 46
justice. Simply put, any potential motivation, consideration, or urge—anything that can move us toward action—fits with this first aspect of a practical reason, the passively experienced motivation.

The second aspect of a practical reason is the active part. It consists in taking up one of those motivations or inclinations and, in a sense, making it our own, endorsing it as a good motivation on which to act. To call a motivation a reason is to imply that it has both parts: an inclination or motivation plus the self-conscious or reflexive endorsement of that motivation as good. In endorsing a motivation, I take it to be a reason to act; and, conversely, to act rationally is to act on a motivation that I take to be right, appropriate, or good—a motivation that I take to be a reason.

5.1.1 Motivation and reasons

This conception of what it means to act for a reason goes a long way, I think, toward resolving the apparent tension between the normative and instrumental aspects of practical rationality. The perception that the normativity or oughtness of reasons is disconnected from people’s real motivations is based on the mistaken idea that the normativity of reasons provides a motivation that is separate from, and rival to, the other motivations the person might have. Thus, my brother reads a chapter draft for me “because it is the right thing to do” rather than to help me out, or a negotiator makes a risky concession “because it is the right thing to do” rather than to secure peace. “Doing the right thing” is conceived as a motivation one might have, to be set alongside other possible motivations, desires, or incentives. And to act for a reason is, on this mistaken view, to be motivated by that normative commitment and not (or not only) in other ways.
The two-part conception of rational motivation shows why this idea is wrong. Normative motivations are not separate from, or external to, the actor’s own motivations. Rather, they are those motivations, actively endorsed by that actor as good. To say, from a third-person point of view, that someone acts for a reason, or (for these are the same thing) acts on principle or acts out of duty, is to say that that person acts with the awareness that the motivation they have is a good one. From a first-person perspective, to be motivated by a reason is, as Christine Korsgaard suggests, “an announcement that a certain way of acting is good.”14 Thus, when my brother reads a chapter draft for me, his reason for doing so is to help me, and he takes that reason to be normatively binding or compelling (which, as I’ve pointed out, is inherent to the concept of a reason). But that does not mean his motivation is now “to do the right thing.” Rather, his motivation is (as it always was) to help me out, and to call this a reason is simply to record the fact that he actively takes this motivation to be good or normatively binding.15

5.1.2 The unity of practical reason

This two-part conception of rational motivation also addresses a second problem: the apparent disconnection between the normativity of reasons and a proper concern for the consequences of actions. To suppose that rationality is about either “doing what’s right” or “achieving a goal” is a mistake for the straightforward reason that goals and desires can themselves be reasons that justify actions. Goal-directed motivations can be reasons in the same way any other motivations can be reasons—they become reasons when the person who acts on those motivations endorses

14 Korsgaard 2008b, 229.

15 I am greatly indebted to Korsgaard (2008b, 228) for this argument and a similar example.
them as good. Thus, if an actor decides that she ought to bring about an outcome $y$, and action $x$ is the best way to achieve $y$, then this instrumental analysis gives her a reason to take action $x$.

But in fact the point can be put more strongly. It is not simply that the actor’s endorsement of the goal $y$ then gives her an “instrumental reason” to take action $x$. Rather, to will the end *just is* to will the means. In other words, effectiveness is built into the very idea of action. Acting is about making ourselves causes, or bringing things about. Robert Brandom points to the symmetry of theoretical reasoning (that is, reasoning about what we ought to believe) and practical reasoning (that is, reasoning about what we ought to do) on this point. While theoretical reasoning is about what we *take to be true*, practical reasoning is about what we *make true*—about effects we choose to have on the world.\(^{16}\) An object of practical reasoning, says Kant, is “the representation of an object as an effect possible through freedom.”\(^{17}\) To act, in other words, is to make real the object of our practical reasoning:

To be an object of practical cognition so understood signifies, therefore, only the relation of the will to the action by which it or its opposite would be made real, and to appraise whether or not something is an object of *pure* practical reason is only to distinguish the possibility or impossibility of *willing* the action by which, if we had the ability to do so (and experience must judge about this), a certain object would be made real.\(^{18}\)

---

\(^{16}\) Brandom 2000, chap. 2.

\(^{17}\) Kant 1997, 50; quoted in Korsgaard 2008b, 68.

\(^{18}\) Kant 1997, 50.
Or, as Korsgaard puts it, summarizing Kant: “To act is to constitute yourself as the cause of an end.”19

So the intuition that rational action must be effective, and must not be “self-defeating,” is integral to the normativity of practical reason, not in tension with it. To act rationally just is to realize a purpose, to put a purpose (a goal the actor has endorsed as good or right) into effect. A rational action is, simply, a good purpose, effectively realized.

That is not to say, of course, that decisions cannot involve choosing between valued goals and considerations that would restrain the pursuit of those goals. Clashes between consequentialist and deontological reasons are real. And indeed, they feature prominently in collective action problems. The problem of opportunistic defection arises because people are subject to incompatible motivations or reasons, some derived from the short-term gains they would receive by defecting, others derived from considerations of longer-term benefits, as well as from commitments to cooperative rules or, perhaps, to identification with social roles prescribing cooperation.

The point, though, is that clashes between norms and desires, or between the “logic of appropriateness” and the “logic of expected consequences,” should be understood as simply conflicts between reasons—or, perhaps, between motivations that purport to be good reasons—and not conflicts between “instrumental” and “normative” rationality. This point tends to be obscured by the fact that we call some reasons “normative” and others “instrumental.” But “normative” as applied to a class of reasons (those deriving from commitments to norms, whether social, moral, or legal) should not be confused with the more basic normativity of

19 Korsgaard 2009, 72. In this formulation, Kant’s notion of autonomy comes in with the idea that you constitute yourself—and not simply a drive, inclination, or motivation within you—as a cause.
practical reason itself. Rationality itself is not “instrumental” or “normative.” In fact, the whole project of identifying rationality with any pre-defined class of principles or considerations is misled. Practical rationality does not consist in the application of a pre-defined set of principles. It is, rather, the activity of adjudicating between the various principles or considerations that present themselves to us, endorsing or rejecting them as reasons for action, and putting those reasons we endorse into effect.20

5.2 Rationality and utility maximization

My goal in the previous section was to show that the apparent incompatibility between the normativity of reasons and instrumental rationality is due in part to misunderstandings about what is implied by the normativity of practical reason. In this section, I will argue that the tension is also due to misunderstandings about instrumental rationality. As I pointed out above, “instrumental rationality” is a misnomer: rationality itself is neither “instrumental” nor “normative” nor anything else. But the instrumental principle of “take the best means to achieve your ends” is often elevated into a conception of rationality itself, and when it is, it contributes to the difficulties we face in reconciling our conflicting intuitions about the normative and instrumental aspects of practical rationality.

5.2.1 The assumption of harmonized ends

What is wrong with conceiving of practical rationality in purely instrumental terms: that is, as the maximization of utility? The key problem is that in doing so we fail to recognize that practical

20 In philosophical terms, this is an argument for a constructivist conception of practical reason, as opposed to what Korsgaard calls “dogmatic rationalism” (2008b, Chaps. 1, 10; Korsgaard 2008c; see also Rawls 2005).
reasoning is often about the choice of ends as well as means. As I argued above, a rational action is a good purpose, effectively realized. In this sense, the rationality of the action is a unity. If we remove the goodness of the purpose or end, leaving only the idea of effectiveness, we have lost something essential to rationality. Reasons are inherently prescriptive; they are “oughts.” But pure instrumentality is prescriptively inert: it cannot tell us what to do because it says nothing about what ends we ought to pursue.\textsuperscript{21} The rationality of actions cannot be assessed apart from the goodness of the purposes for which they are taken, because “rationality” is a judgment that attaches to the action as a whole, including its purpose.\textsuperscript{22}

The idea that rationality is about ends as well as means is, of course, precisely what is denied in adopting an instrumental conception of rationality. That denial might be plausible if we could assume that people’s ends are, from the point of view of practical judgment, unproblematic. What would make a person’s ends unproblematic (at least as far as that person is concerned)? One answer is if there were no conflicts among them. As long as a person’s ends (or purposes, or motivations) were in harmony, there would be no need to choose between them, and thus the question of how that person ought (rationally) to choose between them would not arise.

The notion that a person’s ends are fundamentally in harmony is expressed, in rational choice models, by the idea that individuals have a complete and transitive set of preferences over outcomes.\textsuperscript{23} “Preference” needs to be understood carefully. In modern economic theory the notion of a preference is de-psychologized.\textsuperscript{24} Unlike the earlier utilitarian theories of Jeremy

\textsuperscript{21} Korsgaard 2008b, chap. 1.
\textsuperscript{22} Korsgaard 2009, chap. 7.
\textsuperscript{23} Reflexivity is often given as a third criterion for a preference order, but it is included in completeness, which is defined below. Morrow 1994, 18.
\textsuperscript{24} Hollis and Sugden 1993.
Bentham, John Stuart Mill, and others, the modern doctrine of utility is not meant to refer to actual psychological states like desires or goals. To have a preference for $A$ over $B$ means simply that, given a choice, a person would choose $A$ over $B$. Why she makes this choice, in terms of what psychological needs or desires the choice fulfills, can remain outside the model, “black boxed.”

To say that preferences are complete is to say that between all pairs of outcomes, the actor has a preference: either she would choose $A$ over $B$, $B$ over $A$, or would be indifferent between the two. What completeness excludes is the possibility that there are any outcomes that are simply incomparable. Transitivity implies that for all outcomes, if the actor prefers $A$ over $B$, and $B$ over $C$, she also prefers $A$ over $C$. What transitivity excludes is cycling among preferences. One problem with cycling, and thus a reason to include transitivity in the definition of a preference order, is that cycling preferences are arbitrarily subject to the order in which pairs of outcomes are presented to the actor.

The modern economic concept of utility, finally, is simply a representation of this preference order. Utilities are assigned to outcomes by a utility function, which maps possible outcomes onto a set of numbers, such that for any two outcomes $A$ and $B$, $A$ is assigned a number greater than or equal to the number assigned to $B$ if and only if $A$ is weakly preferred to $B$. Thus, to say that an actor maximizes utility is just another way of saying that she chooses the action whose consequences she prefers to the consequences of any other action.

My contention is that one of the implications of this concept of utility is that it effects a harmonization, or cleaning up, of an actor’s ends in a way that obscures the fact that practical


26 Expected utility applies when uncertainty is included in the model. To get expected utilities, you use a function that maps preferences over gambles, rather than preferences over outcomes, onto a set of numbers.
rationality involves choices between ends, as well as choices between means. That it does so is hidden by the way the concept of utility has been stripped of psychological content. If rational choice models claimed straightforwardly that it is impossible for people to want or be drawn toward different and incompatible ends, they would plainly be wrong as a matter of psychology. But strictly speaking, the models do not have anything to say about such things as desires and wants, and thus seems safe from criticism.

Nonetheless, a good case can be made that the requirement of consistency in choice (given by the criteria of preference completeness and transitivity) amounts, in effect, to assuming the harmony of ends after all. Consider what happens when the consistency of a person’s choices seems to be in question, as in cases of preference reversal. Take, for example, an alcoholic who knows that his drinking is bad for him and for those he loves. Despite this knowledge, he often feels a powerful urge to get drunk. When he is tempted to drink, he knows he should refrain. When the urge to drink is strong he sometimes gives in, and when he does, he experiences genuine regret. But other times he does not give in: he experiences the urge to drink, but bends up his will and overcomes it.

From a rational choice perspective, the problem here is that the alcoholic seems to make choices that are inconsistent. At one moment he chooses to drink, which implies that he prefers getting drunk, and the pleasure or solace it brings, to staying sober. But then when he experiences regret, these preferences are reversed; in the moment of regret, if he were given the choice again, he would choose sobriety.

27 Such preference reversals are often modelled as “hyperbolic discounting.” Ainslie 2001; Heath and Anderson 2010.
The problem is that on a purely instrumental conception of practical rationality, while we might puzzle over the strange preference reversal that has taken place, what we cannot say is that getting drunk was in any way irrational. Within the terms of the model, there is no way to express the alcoholic’s own intuition that practical rationality for him consists in abstaining from getting drunk. Even though getting drunk is, in the alcoholic’s own considered judgment, the wrong thing to do, a purely instrumental model of rationality cannot counsel sobriety. If rationality consists in choosing the best means to your ends, and if, in moments of despair, loneliness, or weakness, the alcoholic chooses the temporary solace and pleasure of getting drunk over staying sober, this can only imply, within a rational choice framework, that being drunk was in fact better to him, at that moment, than staying sober. That preference is revealed in his choice. True, the sudden reversal of preferences suggests that something is amiss, but there is no way of saying what preference is right—no Archimedean point—and thus no way of saying what the alcoholic ought to have done.

It is worth broadening the examples beyond addiction to show that this is not just a special case. Procrastination is another well-known example of practical irrationality. Like giving in to an addiction, procrastinating involves failing to act in accordance with your own judgments about what is the best thing to do. But such failures can happen for any number of reasons, as Christine Korsgaard points out:

You want to ride on this immense roller coaster but you are prevented by terror…[you] lose your nerve and shuffle meekly away…. You want to see a movie but you are too idle to go downtown; you want to go out with him but you

28 Heath and Anderson 2010.
are too shy to call and ask him for a date; you want to work but depression holds you in its smothering embrace. 29

These are all examples in which people act against their own better judgment, examples in which, as soon as people act as they do, they immediately feel regret: the sense that they were not, in that moment, their “best selves,” or that something within them—their “terror, idleness, shyness, or depression”—got the better of them and kept them from doing what they ought to have done. 30 But in all these examples, a purely instrumental conception of practical reason is unable to capture those practical judgments. From a purely instrumental point of view, we can wonder at the sudden reversal of preference, but there is no way to say that one preference or the other was right, no way to say what ought to have been done.

Being unable to make such judgments is deeply unsatisfying. We should hope that any conception of practical rationality, purely instrumental ones included, should do justice to our intuitions in such matters. Furthermore, from an explanatory (as opposed to a practical or evaluative) perspective, sudden and unexplained preference changes threaten to undermine the explanatory value of rational choice models. 31

As a result, when confronted with preference reversals or potentially inconsistent choices, rational choice theorists tend to reframe their analysis in ways that, in effect, assume that people’s ends are fundamentally in harmony, even though, strictly speaking, the models should be agnostic about such things. They do so by changing the outcomes that actors are presumed to have preferences over. Thus, as James Morrow explains, if we are modelling the outcome of an

29 Korsgaard 2008b, 40–41.
30 Korsgaard 2008b, 41.
election, we’d have trouble with the fact that swing voters appear to change their preferences over the outcome during the campaign.\textsuperscript{32} The solution is to make the “outcome” not who wins the election, but instead the \emph{policies} the winner would enact if elected. Then the apparent change in preference is really just a change in strategies or preferences over actions “as voters learn what policies each candidate is likely to adopt if he or she is elected.”\textsuperscript{33} This amounts to removing the apparent preference reversal by “backing up,” moving the level at which we frame preferences to look for what the actor “really” wants. Thus, we suppose that voters don’t care which candidate wins; what they “really” care about is choosing a candidate who will enact their favoured policies. The apparent reversal of preference from one candidate to another was really just the expression of an enduring policy preference, together with a change in belief about what policies each candidate will enact.

How well this “backing up” move reflects the reality of people’s choices will vary. In some cases, the move is entirely plausible, particularly if it involves purely monetary values. In my choice between a fixed or variable-rate mortgage, it’s fair to say that I don’t have any intrinsic preference for one or the other, and that what I really want is to minimize my interest payments over the long run. When it comes to voting, the plausibility of backing up the specification of preferences from candidates to policies is perhaps more debatable.

Be that as it may, the point is that this sort of “backing up” move has the effect of reducing the multiplicity of people’s ends to a harmony, to the thing the actor “really” wants, even though the models claim to remain agnostic about the psychological states of the people they represent. In requiring preferences to remain fixed, as well as complete and transitive,


\textsuperscript{33} Morrow 1994, 20.
rational choice theorists tend to assume—if not in the models themselves then in the way they are applied—that all of an actor’s goals or purposes can really be reduced or harmonized to a common value or scale.

That such an assumption underlies the idea of instrumental rationality is also evident in the language of maximization. The notion of “maximizing utility” has a curious place in rational choice theory. As I pointed out above, technically speaking, the idea of “maximizing utility” adds nothing beyond the idea of “preferring”; utility is a straight mathematical derivation from a preference order. And yet there is a perennial temptation, when constructing formal models, to reverse this conceptual relationship: to suppose that an actor prefers A to B because A “maximizes her utility.”

Despite the fact that it does little or no work in technical terms, the language of maximizing utility is widely used and is frequently given as a shorthand way of explaining what instrumentally rational actors do.

It is instructive, then, that in the term “maximization,” the assumption that all goals or values can be reduced to a common measure lies right on the surface. It only makes sense to speak of “maximizing” if choices can ultimately be related in terms of more and less of something, of some common measure of goodness that all possible goals, desires, or ends share in to a greater or less extent. And as I suggested above, we should not be surprised at this implicit assumption of harmonized ends, because any purely instrumental conception of rationality must go this way. Choices between ends can be removed from the concept of rationality only if, as far as the individual is concerned, no real choices between ends are necessary.

---

34 Morrow (1994, 33–34) points to a number of fallacies that all derive from losing track of the fact that the concept of utility is purely derivative from preferences and has no independent reality or value.
5.2.2 Inner pluralism and rationality as integration

By implicitly assuming a fundamental harmony in people’s ends or goals, rational choice models miss an important element of practical rationality: the fact that rationality involves choosing between ends. Rationality is not only about maximization; it is also about integration—about the integration of conflicting desires and considerations into a coherent course of action.\(^{35}\)

To see this, consider the case of the alcoholic again. From a rational-choice perspective, the alcoholic appears to undergo a strange preference reversal: at one moment he apparently prefers the pleasures of drinking, but at the next moment (or even simultaneously) regrets this choice and would genuinely prefer the goods that come with staying sober. If we want to be able to capture what is amiss in his actions, we might be tempted to make the “backing up” move I described above. We might say that what the alcoholic really wants is not to get drunk, but something drunkenness gets him, such as happiness or utility. And when he experiences regret later, it is because he realizes that he was mistaken: that more happiness or utility can be secured by staying sober. Thus, the rational choice is to stay sober because that is the one that maximizes utility.

But on closer examination, the plausibility of this interpretation falls apart. The alcoholic is not really maximizing. Rather, he is choosing between goods, or perhaps sets of goods: on the one hand, the goods that come with sobriety (say, being less withdrawn from his family and better able to succeed at work) and the goods that come with getting drunk (say, solace or pleasure). The right choice, as he himself acknowledges, is to choose the first. But if he chooses the first, there is no sense in which he is maximizing the goods involved. Rather, he is sacrificing

\(^{35}\) Korsgaard 2008c, 34.
the second set of goods to the first; he is choosing the benefits of sobriety at the expense of the pleasures of getting drunk. Those latter pleasures get to play no part in the outcome he chooses. To the contrary, his greater good in this situation—staying sober—is premised precisely on forgoing the pleasures of intoxication.

The basic problem is that the language of maximization fails to acknowledge the inner pluralism of human experience, the fact that acting frequently involves choosing between competing considerations, divided loyalties, and conflicting values and desires. This is an important bit of psychological realism: it is simply inaccurate to represent human beings as having a single inner voice, a perfectly integrated set of desires and evaluations, such that any outcome could be ranked on a single scale of goodness or utility. It is not the human condition to be handed, as it were, a pre-packaged, unambiguous “good,” such that the question, “what should I do?” can always be answered with, “whatever brings more of that.” Our experience of goodness or value is, simply, more fragmented than that.

That fragmentation of value does not necessarily extend to the fragmentation of the self. This is not an argument for the dissolution of subjectivity. To the contrary, the conception of practical reason I presented above assumes that it is possible for me, myself as a whole, to act: that is precisely what happens when I endorse the motivation or purpose of my action as a reason. The point, rather, is that the integrity of the acting subject is an achievement. As Christine Korsgaard argues, practical reason is about the way we knit the manifold of our desires and urges, some compatible, some incompatible, together into a will, into a coherent conception that we can self-consciously endorse as good:

37 See note 19 above.
A non-human animal finds herself in a teleological world, a world in which things are already marked out for her as her food, her mates, her offspring, her enemies…. But in the human mind, the development of a certain form of self-consciousness—consciousness of the potential grounds of our beliefs and actions—breaks up this teleological conception of the world. It shatters the world into a mass of perceptions or experiences, and the self into a mass of desires and fears and impulses… We are faced with the task of unifying the mass of perception into a conception of the world that enables us to find our way around and act effectively, and of unifying the mass of desires, fears, and responses into a self that can stand behind its movements as their author and so claim them as its actions.  

From this perspective we can see what the language of utility maximization misses. It suggests that the alcoholic’s problem is just that he got his sums wrong, that he made an error of calculation, and so failed to maximize. But that’s not really it. His failure is not so much one of maximization as of integration: a failure, that is, to successfully pull together a coherent course of action out of the incompatible motivations he experiences. His problem, in a sense, is that he remains fragmented, split between warring impulses within him, unable to act as his whole self.  

The idea of utility maximization contributes to the tension between normative and instrumental conceptions of rationality by appearing to offer a way of understanding rationality in purely instrumental terms. But on closer examination, instrumental effectiveness cannot stand

______________

38 Korsgaard 2008c, 38.
alone as a concept of rationality, as a rival to a normative conception of rationality. As I argued above, rationality itself cannot be equated with any particular set of reasons or considerations, whether instrumental or otherwise. Rather, it is the faculty or activity of choosing between reasons, and putting those reasons we endorse into effect. Thus, rationality is best understood as a unity of both instrumental effectiveness and normatively compelling purpose.

5.3 Rationality and collective action problems

The conception of practical rationality I have defended here has important implications for how we understand collective action problems. Collective action problems are often conceived as pitting normative and instrumental logics or forms of rationality against each other. But that view distorts the nature of rational choice in the face of collective action problems. The better route is to understand collective action problems as being simply a special case of the general predicament, faced by all acting subjects, of choosing among the various motivations and purposes that suggest themselves, and putting those motivations or purposes that they endorse into effect.

What is special about collective action problems is the arrangement of conflicting motivations (and thus potential reasons) they involve. We speak of collective action problems when actors are simultaneously motivated to both cooperate with, and defect from, a shared plan or rule. Cooperation is generally urged by the value of goods that can only be achieved by working together, but it can also be urged by commitments to upholding social norms, by social roles or identities, or even, in established cooperative regimes, by force of habit. Defection, on

---

40 See Chapter 2 above.
the other hand, is generally urged by another set of goods: goods that are not common but private, redounding to the actor herself or to her own constituency, but not to the other actors she faces. But like cooperation, defection can also be urged by non-consequentialist considerations such as social norms (though here norms that apply within the actor’s own group rather than between her and her opponents), identities (though, again, more often within-group identities than across-group identities), and force of habit.

To act in such situations, as anywhere, is to adjudicate between these conflicting motivations and purposes, and to put into effect those the actor takes to be best. The way actors are pulled in these two directions is what gives these situations their special character. As I pointed out in Chapter 2, if actors’ interests pushed them straightforwardly toward cooperation, there would be no need for shared rules that are designed to weigh as decisive reasons for cooperation. And if their interests pushed them straightforwardly toward defection, rule-based cooperation would be neither possible nor desirable. Collective action problems present themselves because interests or motivations within each actor are divided.\(^41\)

This understanding of collective action problems contrasts with what I think is the more common understanding, namely, that collective action problems create a conflict between normative and instrumental rationality, or normative and instrumental logics. On the common view, cooperation means following a rule, but this rule is conceived as \textit{external to the actor’s own motivations}. It is a demand or prescription that stands apart from the actor and confronts her: a moral fact, out there in the world. At that point, she can either act “normatively,” in which case she accedes to the normative prescription \textit{instead of} acting on her own interests, or she can act “instrumentally,” in which case she ignores what is normatively right and simply pursues her

\(^{41}\) Cf. Schelling’s (1960, 89) concept of a “mixed-motive game.”
own goals. Thus, as I argued in section 5.1, once the normativity of reasons—in this case the reasons counselling cooperation—is seen as external and rival to the actor’s own motivations, a specious opposition between normative and instrumental “rationalities” is set up.

Notice that this mistaken view of the actor’s situation fits with both the rational choice and sociological research traditions I discussed in Chapter 4. In rational choice approaches, people follow rules when rules and inner motivations happen to align, or if they are brought into alignment by arranging the circumstances and consequences of action so as to incentivize compliance. In other words, the theorist has to “earn” the compliance by showing that it is an equilibrium.42 From a sociological point of view, rule-following depends upon a similar alignment between the rule and the actor’s own motivations, only here the alignment takes place through the actor’s being socialized in the right way.

Either way, when the relationship between rules and action is conceived in this way, what is missing is the idea that the actor might follow the rule because the rule gives her a reason to act. What is missing, in other words, is the activeness of the acting subject who takes the rule to be a good one. As soon as compliance with rules is understood as a matter of aligning a normative demand that is “out there,” external to the actor, with an inner world of motivations, the activeness of choice is lost. If the actor follows the rule because her situation is arranged so that the pursuit of her own ends drives her to do the right thing, in what sense can we say that she has been guided by that rule or reason?43 It wasn’t her: it was the arrangement of incentives that did the trick. Likewise, if the alignment between the rules and her own motivations comes about through socialization—if, in other words, following the rules just is what she wants to do—it is

42 E.g., Landa and Meirowitz 2009, 442.
43 Korsgaard 2008b, 44, 50.
again hard to see how she was guided by those rules. Ultimately, her compliance is due not so much to her as to the social forces that shaped her. Thus, whether the alignment between the moral-rational “fact” about what she ought to do and her own motivations is brought about by incentives or socialization, we cannot really say that in following a cooperative rule she has acted for a reason, or that the cooperative rule motivated her as a reason, that is, as a motivation that moved her by her perception of its rightness.

This is a tricky point, and I want to be clear about it. It is not that coercive incentives and socialization cannot bring about cooperation. To the contrary, as I argued in Chapters 2 and 3, successful systems of cooperative norms can and, where possible, should be supported both by incentives and by the socialized dispositions of the actors involved. The point, rather, is about how these incentives and socialized dispositions (and, for that matter, anything that inclines an actor toward cooperation) should be thought of as motivating. If they are to motivate the actor as reasons, then they should not be thought of as external to the actor, as moral or rational facts “out there,” theoretical knowledge that has to somehow stick to the actors motivationally. Rather, reasons are inherently motivating: in fact, reasons are motivations, and they are the actor’s own motivations. They are, specifically, the motivations that the actor endorses as good, binding, or compelling, and thus puts into effect. Reason-based cooperation occurs when people find the reasons that counsel cooperation to be compelling, and when they endorse those reasons over those counselling defection. In this case, the reason itself explains their cooperation, and there is no need to posit an additional mechanism to explain why people act on the reason.

Finally, the understanding of practical reason I have defended here also has implications for how we think about defection in the face of collective action problems. If we see rules or reasons supporting cooperation as rival to the actor’s own motivations—so that the actor follows
the rule *rather than* pursuing her interests or her own goals—then defection would signify a choice that is amoral, non-normative, or purely instrumental. But if, as I have argued here, we see reasons as being simply the motivations the actor endorses, a different understanding of defection emerges.

When an actor defects, it is not that she is acting according to an instrumental rationality or logic as opposed to a normative one. Rather, there are two possibilities, depending on whether her defection was in accordance with her own practical judgments. If it was, then her defection does not represent an amoral or purely instrumental action, but rather a different judgment about what *is*, in fact, the right thing to do. That is not to say that we, or anyone else, have to agree. The point is that she took her motivation to be a reason, an imperative that was not only motivating, but right. And that is why it is possible for others to judge that action. It is the sort of thing that can be right or wrong precisely because it claims to be a good reason. If it were only a motivation without that normativity—that claim to validity—it could not be judged. It would simply be.

The second possibility is that she did not think defection was the right thing to do, but she nonetheless defected against her own better judgment. In this case, defection is a form of practical irrationality or akrasia: the inability to form a coherent will out of a multitude of warring impulses.\(^44\) This could happen in any number of ways. It could be the result losing her nerve or being afraid. It could be the result of impatience or opportunism, of placing too high a value on short-term gains at the expense of longer term gains. And where the actor involved is a

\(^{44}\) Heath and Anderson 2010.
corporate actor—a state, a rebel group, a caucus—it could be various forms of instability or disorder among factions within.\(^{45}\)

In this second case, we might be tempted to describe defection as the result of instrumental reasoning, particularly if it was the prospect of short-term or selfish gains that motivated the actor to defect in spite of her better judgment. But to describe her actions as the result of instrumental \textit{reasoning} does not work, because the actor herself does not see her choice as reasoned or rational. If you are inordinately motivated by short-term gains, for example, the problem is not that you have skillfully matched means to ends (i.e. that you are instrumentally rational), but rather than you have chosen the wrong ends, the lesser good over the greater one.\(^{46}\)

\textbf{5.4 Conclusion}

If deliberation is to play a role in the emergence of shared rules and collective action, it must be possible for people to be motivated by reasons. Saying how reasons get their motivational grip on people is an important part understanding deliberative will-formation.

In this chapter I have tried to answer that question, paying particular attention to the relation between the normativity of practical reasons and the idea of instrumental rationality. The normative and instrumental aspects of rationality are sometimes thought to stand in tension with each other, as rival concepts of rationality itself, or as rival “logics.” And that tension is particularly high in the face of collective action problems, which appear to pit cooperative norms against self-interest.

\(^{45}\) List and Pettit 2011, chap. 5.

\(^{46}\) Keep in mind that choosing the short-term reward is in this example irrational \textit{ex hypothesi}. One could imagine various scenarios, of course, in which it is not.
But this way of understanding practical rationality is mistaken. First, acting for a reason does not mean acting from a sense of duty or rightness *rather than* other motivations the actor has, including her own interests. It means, instead, that the actor takes her own motivations to be good ones, and is motivated by the perception of their goodness. Second, a concern for the consequences of one’s actions is not external to the normativity of practical reason, but rather inherent to it. Practical rationality is about putting into effect those purposes, considerations, or goals we take to be good.

This conception of practical reason provides an important foundation for thinking clearly about collective action problems. Collective action does not depend on the type of rationality people follow. It depends, rather, on what the actors involved take to be compelling reasons, which involves choosing among goods, as well as choosing the best means to achieve them. From the perspective of collective rationality, this moment of endorsement or judgment is where the action really lies. Once an actor judges reasons prescribing cooperation to be compelling, the fact that she then acts on those reasons requires no further explanation, because reasons are inherently motivating. In the next chapter, I will look at the kind of reasons that can ground collective actions: public reasons.
Chapter 6    Public Reasons and Collective Actions

The previous chapter ended with a picture of what practical rationality implies in the face of collective action problems. To act rationally in such situations (as anywhere) is to adjudicate among the various purposes or inclinations that present themselves to an actor, and to put into effect those that the actor takes to be best. In collective action problems, people’s motivations and purposes are divided—they are pulled toward both cooperation and defection. The rationality of cooperation depends ultimately upon whether they find reasons prescribing cooperation to be compelling or decisive.

So far, so good. But why should people find cooperative reasons compelling, rather than non-cooperative ones? What makes cooperation the rational choice (when it is the rational choice)? In order to answer these questions, we have to go beyond the individual-level account of practical rationality developed in the previous chapter. In situations of strategic interdependence, rational choices also depend upon others. Collective action is irreducibly interactive; it is something that people do together. To explain the emergence of cooperation in anarchy we thus need an appropriately interactive account of what makes cooperative reasons compelling in the face of collective action problems.

The answer, I will argue, lies in the notion of public reasons. Public reasons are reasons that are shared, but shared in a particular way. Reasons are public when they are addressed to each other as justifications. The validity of public reasons depends, however, upon the existence of a public, which is a group of people who consider themselves accountable to each other and whose reasons for acting are responsive to shared standards of justification.
As I will argue below, in collective action problems, decisive reasons in favour of cooperation tend to be public reasons. That means that whether people cooperate turns on whether they see themselves as forming a public—that is, whether they and the others with whom they are interacting see themselves as being accountable to each other. In this way, the challenge of collective action is ultimately the challenge of constituting a public.

In later chapters, I will argue that deliberation plays a crucial role in the emergence of public reasons and the relationships of mutual accountability that sustain them. Thus, by identifying the relationship between public reasons and collective action, this chapter takes a key step toward an account of how deliberation shapes cooperation in anarchy.

The chapter is organized as follows. In section 6.1, I will explain what a public reason is. What makes a reason public is not its content but its “justificatory scope.” A public reason is addressed as a reason to everyone affected by a given collective action problem. By contrast, I define a reason as “private” if it is meant to justify actions only to part of those people who are affected by collective action or its absence.

In section 6.2, I argue that mutual accountability is the essence of publicity. What makes a reason public is not that it happens to be endorsed by others, but rather that the speaker sees himself as accountable to others for its rightness. It is the practice of mutual accountability that creates the “space of reasons” within which others can either agree or disagree with a claim. The mere congruence or overlap of reasons is neither sufficient nor necessary for them to be public.

In section 6.3, I elaborate the implications of public reasons for collective action. In collective action problems, the decisive reasons for cooperation are typically public reasons. I argue that interaction based on public reasons differs fundamentally from the sort of instrumental interactions represented in game theoretical models. Actions based on public reasons are, in
effect, shaped by mutual accountability at the level of the actors’ preferences themselves. The result is that public reasons can ground cooperative actions that would not be taken if actors treated all reasons as being devoid of intersubjective accountability, as game theory, in effect, assumes they do.

Section 6.4 argues that the existence or non-existence of a public is an essential parameter of individual decision-making in contexts of strategic interdependence. Facing a collective action problem, the answer to the question, “what should I do?” depends on whether, and to what degree, people see themselves as standing in relations of mutual accountability to others.

Finally, in section 6.5, I show that collective action does, in fact, depend upon mutual accountability between actors, and not merely on actors’ behavioural expectations of each other. This result contradicts the assumptions of both rational choice and normatively prescriptive understandings of collective action.

6.1 The idea of a public reason

What is a public reason? There are two main senses of publicity. The first is perhaps the most straightforward. A public reason is simply a reason given in public, a reason made accessible to a wide audience. Publicity, in this sense, is the opposite of secrecy. If public reasons are openly given, private reasons are given to select groups, behind closed doors—or they may even be reasons an actor acknowledges only to herself.

Publicity in the second sense refers not to how widely a reason is disseminated, but rather to a property of the reason itself. Public reasons, in this sense, are reasons that belong to, or are in one way or another acceptable to, everyone in a political community and not only to particular
members of that community. In other words, public reasons have to be part of an “overlapping consensus,” or what Rawls calls a “political” conception of justice, that does not depend for its justification on any particular set of religious or philosophical beliefs. Reasons that are public in this sense are not necessarily public in the first sense. Thus, for example, a judge who conducts a secret review of a government request for a phone tap warrant will necessarily keep the reasons for her decision private in the first sense. But we should still hope that she decides for reasons that are public in the second sense.

In this chapter, I will adopt a notion of publicity that is closely related to the second sense. Publicity, on this view, is a matter of the justificatory reach of a reason, that is, of the range of people for whom the reason is meant to be binding or compelling. Reasons are public, in other words, by virtue of their being addressed to a public as reasons. That is different from saying simply that they are heard by a public. What matters is that they are presented to the public as valid claims, as claims that have the normativity or “oughtness” of reasons.

That raises the question, who is the public? The public, as I will use the term here, is the set of actors involved in a given collective action problem. The term “collective action problem” identifies a certain type of interdependence or mutual affectedness among a group of people. To recall from Chapter 2, collective action problems occur where people have (a) conflicting

---


2 Rawls 2005.

3 These two senses of publicity are related, though, in that one reason we should want government decisions to be public in the first, transparency sense is as a guarantee that they are public in the second, substantive sense.

4 Korsgaard 2009, 192–95.

5 More precisely, the people who are subject to a collective action problem have the potential to be a public. As I will argue later, though, whether they actually form a public depends on whether they see themselves as accountable to each other in terms of reasons.
interests, (b) a common interest in maintaining a cooperative scheme, and (c) incentives to defect from that scheme— incentives that, if followed, lead to an inferior outcome for all. This kind of interdependence is problematic because the people involved in it often become trapped in patterns of mutually destructive competition even though they could be cooperating in ways that would greatly benefit all of them.

A public reason, then, is addressed as a reason to everyone affected by a given collective action problem. In this sense, public reasons have a maximum justificatory scope. They are addressed as reasons or valid claims to everyone who might be affected by them. Other reasons, however, have a more restricted justificatory scope. These are reasons that an actor addresses as reasons only to part of those people who are affected by a collective action problem. I will refer to such reasons as “private” or “partial.”

To illustrate this distinction, consider some of the practical reasons at work among the parties to the conflict in Northern Ireland. Among unionists, for example, defending the union with Great Britain was a central goal, and the reason for many of the actions they took. And some unionists, most prominently Ian Paisley, were also strongly motivated by anti-Catholicism and the desire to maintain the identity of Northern Ireland as a Protestant state. As far as the interactions between unionists and nationalists are concerned, reasons like these are private. When unionists cite the need to defend the Union or, in a more religious vein, to fight Catholicism, their reasons are meant to justify their actions to other unionists and Protestants, not to nationalists or Catholics, who can hardly be expected to find such claims persuasive.

---

6 This definition of publicity reflects the democratic concern with including “all affected interests” in part but not in whole (Whelan 1983; Goodin 2008, chap. 7). There may be some forms of problematic affectedness that are not captured in my definition of a collective action problem (cf. Bohman 2007, chap. 1). Interactions that are purely exploitative—in which all the benefits of cooperation go to one party—do not give rise to collective action problems as I have defined them here. That only underscores the argument I made in Chapter 2, that while overcoming collective action problems is immensely important, it is by no means the only normative concern we should have.
By contrast, some of the reasons and justifications cited by both unionists and nationalists were public in that they were offered by leaders not only to their own constituencies but directed also across the lines of the conflict. In the 1990s, securing peace became an important public reason, as leaders from both communities responded to the growing public desire for peace and a broad-based public rejection of political violence.\(^7\) Likewise, ideals of equality, mutual respect, and “parity of esteem” between the two communities provided justifications that crossed partisan lines.\(^8\) What made these reasons public is that they could be addressed as reasons, as valid claims, to a public that included both sides of the conflict, and not only to members of one group or the other.

It is important to notice that what makes these reasons public or private is not whether they are stated openly or behind closed doors, but rather the audience to whom they are addressed as reasons. Thus, a unionist might openly, or “publicly,” proclaim the need to fight Catholicism, and even do so, defiantly, to Catholics themselves, but his anti-Catholicism remains a private reason as I will use that term here. It is private because while it might be heard by, or even proclaimed to, others outside his group, it is not offered to others as a justification that they might be expected to find compelling.\(^9\)

Categorizing reasons as “public” and “private” with reference to collective action problems, as I have here, embodies a normative judgment. In the context of a collective action problem, there is something wrong, or at least problematic, with private reasons. That is not necessarily true of all reasons with a limited justificatory scope. In many cases, it is quite

\(^7\) Mulholland 2002, chap. 5.
\(^8\) O’Flynn 2006; O’Neill 2003.
\(^9\) The terms “public reason” and “private reason” are used in the same way by Mitzen (2013, 53–54) and Korsgaard (2009, chap. 9).
legitimate to act on reasons that are valid only within a particular group, such as a family, a group of friends, a club, association, church, or nation. For example, any group of people who share a particular goal (say, a soccer team trying to win a game) will share practical reasons related to that goal that are not reasons for others. But in the context of a collective action problem, reasons with a limited justificatory scope, that is, “private” reasons, take on a character of *partiality*. As I will argue later, their partiality—the fact that they fail to address everyone involved in the collective action problem—often contributes to the tragic outcomes that unresolved collective action problems create.

### 6.2 Reasons and accountability

The definition of a public reason I have put forward, however, hides some rather tricky conceptual problems. A public reason is one that is addressed *as* a reason to all those affected by a given collective action problem. But how do we tell if a reason meets this criterion?

I will argue that what makes reasons public is ultimately a particular attitude people take toward one another, specifically, the attitude of holding one another mutually accountable for the validity of the reasons they espouse and act upon. This relationship of mutual accountability is what I mean by saying that a public reason is addressed to another person *as* a reason. To offer a reason to another person *as* a reason—that is, as a normatively binding or compelling claim—is to make yourself accountable to that person for the validity of that claim.\(^\text{10}\) The act of offering a reason in effect says, “I take this to be a good reason, and I hereby commit myself to its goodness.” That implies an obligation to retract your statement or change your position, and

\(^\text{10}\) Brandom 1994.
perhaps your actions, if it is shown not to be a good reason. And it implies an expectation on the speaker’s part that the listener will accept the reason. Reasons always make claims on those they are addressed to: they impose oughts—they say what you ought to do or believe—conditional on their validity. In short, the act of addressing a reason to another implies mutual accountability. It is ultimately this mutual accountability that distinguishes public reasons from private ones: public reasons imply and affirm relationships of mutual accountability between the parties to a conflict, whereas private ones do not.

The connection between public reasons and mutual accountability is not immediately obvious. Public reasons are often conceived in a non-relational and individualistic way. On this view, what makes a reason public is simply the fact that it is acceptable to others. Public reasons are simply reasons that overlap. For example, states in a conflict might observe a rule prohibiting certain kinds of actions, such as laying anti-personnel land mines. Each state takes the rule to be binding; each accepts it, that is, as a good reason not to lay mines. And because they both accept that rule, we can think of it as a public reason. There is no need, on this view, to talk about mutual accountability or the way those states relate to each other. It is enough to say that each state is governed by the rule, and that the rule is the same for both.

But on closer examination this conception of publicity doesn’t hold up. The problem is that it gets the ontology wrong. It takes reasons to be the reality, and the relationship between the actors to be an epiphenomenon: simply a way of talking about the properties of the reasons they hold. There are reasons (a reason being a motivation or principle that the actor endorses as good or right); some of those reasons happen to be shared; and if so, we call them public reasons. And a “public” (if there is such a thing) is simply an incidental property of those reasons: it is simply the set of people who happen to endorse those reasons at any particular time.
In fact, though, the public, conceived as a group of actors who relate to each other on a basis of mutual accountability, is a condition for the existence of public reasons. As I argued in Chapter 3, it only makes sense to think of a public reason as a reason—that is, as a valid claim—in the context of a social practice of attributing and undertaking normative commitments.\(^\text{11}\) The public is in many ways a spatial concept.\(^\text{12}\) A public creates a “space of reasons”\(^\text{13}\) like a public square: a space in which the claims and counter-claims we call reasons come to have meaning because of the way they function in a social practice of attributing responsibilities to ourselves and others. These relational attitudes—the attitude of holding others responsible, and holding yourself to be responsible to them—underpin and give meaning to the thing we call a “reason.”

### 6.2.1 Overlapping reasons are not sufficient for publicity

The relational character of publicity is also demonstrated in the fact that the mere overlap of reasons is neither sufficient nor necessary for reasons to be public.\(^\text{14}\) First, the fact the reasons overlap is not sufficient to make them public. Reasons that just happen to be shared, and that just as well might not have been shared, are not public. A public reason is endorsed as public, as a reason not just for me but for everyone. In endorsing a reason as public, I am making a claim on the other members of the public that they should also endorse this reason. If I could say, “this is a reason for me, but it doesn’t have to be for you,” it is not a public reason, because public reasons

\(^\text{11}\) Brandom 1994.


\(^\text{13}\) Brandom 1994, 5; Erman 2010.

\(^\text{14}\) As Chambers (2009b, 354) argues, “Public reason is constructed not found. This is a feature of all modern conceptions of public reason...” (including those of Hobbes, Locke, and Kant). That is, public reason is not simply reasons that happen to be common or shared. Rather, the notion of public reason is “constructivist” in that it embodies particular political assumptions or ideals.
are defined by their normative or justificatory reach. It is this implicit normative claim, and the relationship of mutual accountability it implies, that makes public reasons relational or social in a way that private reasons are not.

To see why accountability is essential to publicity, it is helpful to examine two types of interaction in which people act on similar reasons but do not address, or think of, those reasons as intersubjective claims. First consider interaction that is instrumental but non-strategic. Here, the consequences of each person’s choice depend on what others do, but her choices do not. Competition in a perfect market is an example of such action. Wheat farmers competing with each other will each try to maximize their profits by pricing their wheat just below the current market price so as to sell more. But since all the farmers do the same thing, the price of wheat falls and they all end up making less than they were originally. In this case, each farmer happens to adopt the same practical reasoning: each thinks, “to maximize my profits, I should price my wheat just below the current market price.” But this is a partial or private reason, not a public one, because each farmer in effect treats it as a reason only for himself, and not as something the others should also do. (And indeed to prescribe this as a reason to others would be counterproductive.)

Now consider a trickier comparison: a case of strategic interaction. In strategic interaction, not only the consequences of each person’s choice, but also those choices themselves, depend on what each person thinks others will do. An example of such interaction might be coordination on the DVD video format between companies that produce DVD players and those that produce DVDs. Once this particular format for video is in place, the actors

---

15 This move from non-strategic to strategic interaction is a result of moving away from the conditions of competition in a perfect market. Milner 1998, 770.
involved in effect act upon the same practical reasoning—namely, that DVD is the right format to use for both video playing machines and movie releases. But now, unlike the wheat farmers, they endorse this reasoning because they know that others endorse it, too.

We now have two actors acting on the same reason, a reason, moreover, that they adopt because they know it is shared. So is this—the DVD video format—a public reason? No, it is not. What keeps it from being a public reason is that the actors lack the mutual accountability necessary for publicity. In strategic coordination, the players do not make claims on each other, the sort of claims that are inherent in the essential “oughtness” of reasons. For a reason to be public, its “ought” must extend between the actors involved. But in strategic coordination, the players do not use reasons to make claims about what each other should do. While they do notice and respond to each other’s reasons—that is what strategic reasoning is all about: figuring out what the other is thinking and responding accordingly—this responsiveness does not have a normative character. It is responsiveness but not accountability. That accountability is missing because they are not addressing reasons to each other as reasons.

This becomes clear if we consider what would happen were coordination to break down. In the example above, suppose that the media company started releasing movies in a new proprietary format instead of DVD. Both companies are harmed. The media company doesn’t sell many movies because the video-machine company doesn’t make machines that can play that format. And the video-machine company is harmed because popular movies are no longer playable on their machines.

This is manifestly an irrational choice on the part of the media company. And “irrational” is, as always, an evaluative judgment. The media company ought not to have switched to a non-
standard format. And given the harm it did to the company, the people who made that decision have to answer for it.

But to whom are they answerable? To people who are essentially part of that same company: to other employees who counted on the leaders of the company to make good business decisions (and who might even lose their jobs if the company flounders) and to shareholders whose shares have now lost value. Notice, however, that they are not answerable to the company that makes video machines, even though that company was also harmed by their bad choice. While the video-machine company will surely be upset that its profits, too, have gone down, it does not have a grievance—it cannot claim to have been wronged—because by definition companies who are interacting strategically do not have obligations to each other in the first place. In purely strategic interactions, the people involved do not see themselves as being accountable to each other. Thus, even though they may coordinate on similar reasons or rules, those reasons do not function as public reasons.16

6.2.2 Overlapping reasons are not necessary for publicity

I have been arguing that the mere overlap of reasons is not sufficient to make them public reasons. But it is also true that the overlap of reasons is not necessary for them to be public. Reasons can be public even if they are not accepted by those to whom they are addressed.17 A public reason requires a public, understood as a group of people who hold each other mutually accountable for the validity of the reasons they give each other, but it does not require

16 This is part of what is captured in Rawls’s (2005) distinction between a “mere modus vivendi” (147) and cooperation that is stable “for the right reasons” (390).

17 Cf. Bohman and Richardson 2009.
agreement. Rather, the publicity of reasons is what makes it possible for people to either agree or disagree about what is a good reason.\textsuperscript{18} Both agreement and disagreement are compatible with the practice of mutual accountability, of addressing reasons to each other, and it is that practice that makes reasons public.

Consider, for example, a trade dispute in which country \textit{A} accuses country \textit{B} of “dumping” (that is, of exporting a product at distorted, below-marked prices) and country \textit{B} denies it. The fact that they have something to argue about shows that the reasons they endorse and act on are not the same. What one considers dumping, and thus illegitimate, the other considers a legitimate result of its comparative advantage in making that product. But despite this disagreement, the reasons they address to each other are still likely to be public. They are still reasons addressed to each other as reasons, as valid claims that (each thinks) the other ought to recognize as valid.

The relationship of mutual accountability implied in public reasons must be compatible with disagreement about what, in fact, are good reasons. Otherwise, deliberation and argument would be either unnecessary or impossible. If offering public reasons meant that those reasons had to already be endorsed by listeners, then arguing would be unnecessary. If, on the other hand, agreement is necessary for reasons to be meaningfully addressed to each other as reasons, and the hearers do not already agree with the reasons, meaningful argument would be impossible. For deliberative influence to be possible, it must be possible to offer reasons that are public—and thus meaningful as reasons, as valid claims, to others—without those reasons being part of a prior consensus.

\textsuperscript{18} Erman 2009.
As a matter of practice, distinguishing between public and private reasons can be difficult. Suppose, for example, that an Israeli settler in the West Bank tries to justify his land claim to a Palestinian by pointing out that God gave the Land of Israel to the Jewish people. Is that a public reason? On the one hand, it clearly seems not to be. That claim cannot hope to be taken seriously by Palestinians—neither by devout Muslims whose religious tradition would deny the claim outright, nor by secular Palestinians who view all religious justifications of territorial claims as manifestly invalid. On a Rawlsian conception of public reason, “God gave us the land” is a paradigmatic case of a non-public reason, a reason that is clearly not part of an overlapping consensus or a political conception of justice that stands free from people’s particular religious or philosophical “comprehensive doctrines.”

But on the other hand, perhaps it is possible to imagine a devout Jewish settler genuinely offering this reason to a Palestinian as a reason—and that, according to the definition I set out above, would seem to make it public. After all, to that Jewish settler, God’s laws and promises are good reasons, even the paradigm of a good reason. Moreover, such a person might think that other people should be able to recognize these as good reasons too. I can imagine the settler saying, “I mean that seriously. God really did give this land to us. If God had given this land to you, then I would accept that, but that is not what God did.”

Could this be a public reason? I am still inclined to say no, but more important are the grounds on which that judgment is based. The judgment ultimately comes down to whether the speaker is really making himself accountable to the other or not. What makes the publicity of this reason suspect is the fact that in offering it the settler is not really making a criticizable claim; he

19 Rawls 2005.

20 My thanks to Stewart Prest for helpful comments on this point.
is not really open to being convinced were the Palestinian able to show that he is wrong. It is not really a matter of the settler’s certainty; it is not that the settler thinks his claim right. The problem, rather, is that in citing this reason he is not really making himself accountable to his interlocutor for its rightness. In offering a reason like this one—plainly rooted in a religious tradition that he knows the other does not share—he is not really involving himself in the practice of building shared standards of judgment and normative accountability. And it is that practice that is the ultimate ground of publicity.  

6.3 Public reasons and collective action

This conception of public reasons is crucial to understanding the potential role of deliberation in structuring cooperation, especially in the face of collective action problems. As I noted in Chapter 5, collective action problems are distinguished by the way they pit reasons for cooperation and for defection against each other. People facing a collective action problem are simultaneously motivated to both cooperate with, and defect from, a shared plan or rule. But with the distinction between public and private reasons in hand, we can now be more specific. What is special about a collective action problem is not simply that it pits reasons favouring cooperation and defection against each other, but more specifically, that it pits public and private reasons against each other. A collective action problem, then, is a situation in which people have compelling public reasons to cooperate that are incompatible with at least some potential private, or partial, reasons.  

To see how this looks, consider a well-known collective action problem in international relations: contributions to common defense. In a defensive alliance like the North Atlantic Treaty Organization (NATO), all the members benefit from the defense capabilities and deterrent effect
of their combined strength, but individual member states would benefit most if they could reduce
their own military spending while still maintaining the protection of all the others. If each state
tries to realize this maximum benefit, however, the result is that the alliance as a whole is
weakened—possibly to the point where it can no longer serve as an effective deterrent. Common
defense, in other words, is a classic public good, and creates the same incentives to free-ride as
other jointly produced public goods.\textsuperscript{21}

Notice that in this situation, reasons weighing decisively in favour of cooperation are
public reasons. There may be many such reasons, including the shared security that comes with
belonging to a strong defensive pact, the fact that contributing to NATO is an important part of
many member states’ identities,\textsuperscript{22} and the fact that NATO members have made formal
commitments to maintain a particular level of defense spending. These reasons are public in that
NATO members can address them to each other as reasons, that is, as justifications claiming
validity. Reasons like these are, in effect, points of accountability between members. Thus, state
$A$ can point out to state $B$ that in shirking on defense spending it has violated a commitment it
made and weakened their common defense, and $B$ would owe $A$ a response. (The United States
often makes exactly this complaint about smaller European countries failing to bear a fair share
of the burden of maintaining NATO’s capabilities.) By contrast, the reasons member states might
have to free-ride in the alliance tend to be private, partial reasons, such as the desire to minimize
its own government spending.

\textsuperscript{21} It is perhaps more accurate to call this a “club good” rather than a public good, since it is available only to
countries within NATO, rather than being genuinely non-excludable, as public goods are. However, once states are
in the club (and as long as there is no serious chance of their being kicked out), common defense functions like a
public good. For the distinction between public goods and club goods see Heath (2009b, 82). The distinction was
introduced by Buchanan (1965).

\textsuperscript{22} Risse 1996.
If reasons in favour of cooperation tend to be public reasons, then whether a group succeeds in acting together will have much to do with whether they come to see themselves as a public. That is to say, collective action turns in large measure on the extent to which the people involved see themselves as accountable to each other for the rightness of the reasons they act on, and thus consider (cooperative) public reasons to be compelling. The capacity of a group of interdependent actors to engage in collective action grows as they expand their justificatory communities beyond themselves and their own constituencies to include all those affected by the common problem.

6.3.1 Comparison to game theory

In order to clarify what is at stake in this argument, it is helpful to contrast it to game theoretical treatments of collective action problems. The notion of acting on public reasons I have presented here departs from game-theoretical assumptions in fundamental ways. The key departure is that publicity and mutual accountability in practical reasons cannot be accommodated within rational choice models. That, in turn, makes a difference for the kinds of actions we should expect. As I will argue below, public reasons make possible forms of cooperation that cannot be understood in rational-choice terms.

The first difference is that game theory necessarily represents practical reasons as being private, not public. That may not be immediately clear, since the language of public and private reasons is not used in game theory. But the basic assumptions of game theory in effect treat all reasons as private reasons. That assumption is rooted in the instrumental conception of practical reasons as being private, not public. That may not be immediately clear, since the language of public and private reasons is not used in game theory. But the basic assumptions of game theory in effect treat all reasons as private reasons. That assumption is rooted in the instrumental conception of practical

---

23 Note that public and private reasons are not the same as public and private information. See Chapter 7.
rationality that such models assume. Rationality here is understood as means-ends or instrumental rationality. To act rationally is to choose the means that achieve your ends, i.e. to maximize utility. Ends themselves, however, are neither rational nor irrational; they simply are. And that means that they cannot be open to criticism or questioning as to their appropriateness by the other players in the game. In this, the instrumental conception of rationality stands in contrast to the one I defended in Chapter 5.

In game theory, this instrumental conception of rationality is manifest in the fact that preferences are assumed to be fixed. In standard game theory, preferences themselves are not in play; rather, they are part of the background against which players choose the best course of action. As long as preferences are assumed to be fixed, there is no way to see the players as being accountable to each other for the goals they pursue. Preferences are points of fundamental unaccountability to others—points of “consumer sovereignty.” Thus, each player’s reasons for acting are, in effect, private, not public.

This becomes clearer if we consider what public reasons would imply in game theoretical terms. When people come to be motivated by public reasons, as opposed to only partial ones, the way they interact changes fundamentally. Within a public, what counts as a good reason to act, from the individual’s point of view, comes to depend on what can be justified according to shared standards of judgment. Public reasons imply that people are open to the revision of their goals and purposes through rational criticism and questioning.

If that is true, the players’ preferences cannot be considered fixed when they begin interacting with each other. The very point of public reasons is that actors are responsive not only to each other’s means-ends reasoning—as they are in standard game theory—but also to each other’s judgments about what ends to pursue. This makes interaction that involves public
reasons more deeply “social” than interaction that involves only private reasons. One might say that where reasons are public, interaction between the players is not restricted to influencing each other’s strategies; it also invades the players’ preferences themselves. That might sound impossible—and within the assumptions of standard game theory, it is. But it is not so strange: it is simply what it means to buy into the legitimacy of social rules and commitments, and to consider yourself to be answerable to others for the goals you pursue.

6.3.2 Public reasons and strategic interaction

If public reasons mark a fundamental departure from the assumptions of game theory, we should not be surprised if they also make possible different results from what game theory expects. More specifically, when people act on public reasons, we may see them taking actions that are not part of a Nash equilibrium. That means that communication through public reasons is not simply a coordination device: a way of choosing which equilibrium strategy to play and a convenient answer to the multiple equilibrium problem. Rather, public reasons hold out the possibility of outcomes that would be out of reach if the players treated their reasons and those of others as private.

Perhaps the way to confront the issue most directly is by looking at the standard, one-shot Prisoners’ Dilemma (Figure 6.1). In this situation, the game theoretical analysis is unambiguous: there is a single Nash equilibrium in which both players defect (D,D). (In the repeated Prisoners’ Dilemma, by contrast, there are a great many Nash equilibria, including some in which the


25 For this and related equilibrium concepts, see, e.g. Osborne and Rubinstein 1994.

26 Heath 2001a.
players cooperate all or most of the time.) The tragedy of the Prisoners’ Dilemma, of course, is that mutual defection is inferior to mutual cooperation for both players. Thus, the “rational” choice of defection makes both players worse off than the “irrational” choice of cooperation.

This should, I think, strike us as a strange and counter-intuitive result.\(^ {27}\) (And I suspect that it usually does strike most students as counter-intuitive when they first encounter it.) Why isn’t it rational for the players to move from (D,D) to (C,C), given that it would make both of them better off? The answer, I suggest, has to do with the unilateralism that is built into the game structure. Mutual defection is a Nash equilibrium because in it neither player can improve his outcome by \textit{unilaterally} changing his strategy. Even though each player prefers the outcome of the strategy profile (C,C) to that of (D,D), it is not up to either player to choose a strategy profile.\(^ {28}\) Each player can only choose his own strategy or action, and not the other’s. In the standard 2×2 representation of the game, Player 1 chooses whether they go up or down and Player 2 chooses whether they go right or left, but what neither player can do is choose on the diagonal.\(^ {29}\)

\(^{27}\) Sen 1977.

\(^{28}\) A strategy profile is a set of strategies, one for each player in the game.

\(^{29}\) Chapman 2006, 84ff.
Figure 6.1. The Prisoners’ Dilemma

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3,3</td>
<td>4,1</td>
</tr>
<tr>
<td>D</td>
<td>1,4</td>
<td>2,2</td>
</tr>
</tbody>
</table>

Each player chooses to cooperate (C) or defect (D). Payoffs are represented as utilities.

But when public reasons are in play, choosing on the diagonal is something the players can do—or, at least, public reasons make possible forms of collective action that amount to choosing on the diagonal in game theoretical terms. The players do so precisely by overcoming the strict unilateralism built into game theory.\(^{30}\) This is the game theoretical significance of the sort of mutual accountability implied in public reasons. To be genuinely accountable to someone else means ceding some control over your choices, allowing the other some control over your arbitrary will, and accepting some loss of “consumer sovereignty”—the fundamental unaccountability of preferences. Thus, in a context of mutual accountability, each person’s actions cannot be thought of as wholly unilateral. Each person grants the other some “say” in his own choices, and when that happens, it becomes possible, in effect, to choose on the diagonal: to choose together a cooperative outcome based on a shared, public reason.

\(^{30}\) This idea has been explored under the rubric of “team reasoning” or the shift from an “I” to a “we” perspective in playing games. Gilbert 1989; Sugden 1993; Chapman 2003; Gold and Sugden 2007.
6.3.3 Instrumental and normativized interactions

In light of the profound difference public reasons make to the way people interact, I would suggest that it makes sense to speak of different kinds of interaction, depending on whether the actors involved act only on private reasons, or on public reasons as well. Interaction based only on private reasons is instrumental. In this sort of interaction, people relate to each other only as instruments or tools; each uses the other as a means to his ends by attempting to choose a strategy that affects the other’s payoffs, and thus his choices.\(^{31}\) By contrast, to the extent that people act on public reasons, their interactions become normativized, because in espousing public reasons, they make themselves accountable for the rightness or goodness of those reasons to each other.

To adopt this terminology is to risk misunderstanding, so I want to be clear. Instrumental and normativized interaction should not be confused with instrumental and normative action. The distinction I am drawing here is about the way actors relate to each other (i.e. the way they interact) and not about rationality or logic each follows in acting. On the level of action, the conclusions of the last chapter stand: practical reason is neither “instrumental” nor “normative.” Rather, all practical reasons, regardless of their particular content, are both normative and instrumental: they are normative in their inherent prescriptivity and instrumental in that they are concerned with effectiveness. A rational action is simply a good purpose, effectively realized. One of the burdens of the last chapter was precisely to resist the division of practical reason into different “logics” or different ways of acting.

---

\(^{31}\) That is, they treat each other only as means. Kant 1993b, 36 (Ak. 429). Korsgaard 2009, 193–95.
However, when we shift the level of analysis from models of individual decision-making to models of interaction, it does make sense to distinguish between relatively instrumental and relatively normativized interactions. The difference is an important one: it is the difference between interactions that are devoid of mutual accountability (instrumental interactions) and interactions that are not (normativized interactions).

This means that normativized and instrumental interactions will not necessarily correspond to reasons that are “normative” or “instrumental” in content. This is just an implication of the formality of the concept of a public reason. What makes reasons public or private is their justificatory reach and not their content. In terms of content, both public and private reasons come in all shapes. Public reasons can, of course, arise from norms that actors recognize as obligations to each other. But they can also be reasons of instrumental efficiency, as when a group of actors recognizes a common good and argues about the best course of action to achieve it. In the same way, while private reasons can, of course, be instrumental or goal-oriented in content, they can also refer to moral, legal, or social norms—except that, as private reasons, those norms will be ones that are cited only by a subset of those actors who are involved in the relevant collective action problem. Thus, it is important not to confuse the normative or instrumental character of interactions between players with the content of the reasons they employ.

As a matter of fact, though, in the context of collective action problems, we can point to some contingent connections between normative interaction and reasons that are normative in content. Given the characteristics of collective action problems, the public reasons that opponents adopt in this context tend to take the form of binding norms or rules. Collective action

problems tend to pit private and public reasons against each other: private goals or rationales counsel actions that are incompatible with public goals or rationales.\textsuperscript{33} For this reason, the public reasons that govern action in such situations tend to have the character of trumps: they include, as part of their conceptual content, the stipulation that in cases of incompatibility they should override private reasons, and not vice versa.\textsuperscript{34}

Rules that are binding in this way make it more difficult for an actor to break the rule while maintaining the relationship of mutual accountability that public reasons depend on. Thus, when an actor breaks such a rule, she cannot say to others such things as, “Oh, don’t worry: I respect our shared rules. It’s just that in this case I could gain more by breaking the rule.”\textsuperscript{35} Respecting such rules excludes, as part of their conceptual content, breaking them in order (say) to make more money. Binding public reasons, in other words, anticipate and overrule competing private incentives to defect. That is what it means to say they are binding. Given the incentive problems inherent in such situations, there is a functional rationale for public reasons to take the shape of binding norms.

But it should be remembered that the connection between normativized interactions and reasons that are normative in content (in that they derive from the existence of social or legal norms) is contingent, not necessary. That is one thing that distinguishes the understanding of collective action I am advancing here from more straightforwardly moral theories. The conflict between self-interest and the common good is one of the oldest themes in political thought.\textsuperscript{36}

\begin{flushright}
\textsuperscript{33} This, I think, expresses the problem better than saying that self-interest is pitted against the common good. The problem is not that private reasons are “selfish” (because sometimes they are not) but that they are partial.
\end{flushright}

\begin{flushright}
\textsuperscript{34} Chapman 2005, 121ff; Chayes and Chayes 1993.
\end{flushright}

\begin{flushright}
\textsuperscript{35} Cf. Brandom 1999.
\end{flushright}

\begin{flushright}
\textsuperscript{36} Mansbridge 2001.
\end{flushright}
One venerable approach to this problem is to cast it in moral terms; overcoming collective action problems becomes a matter of doing the right thing. The theory I am defending here, though, shifts the emphasis away from moral prescription based on the content of reasons to a consideration of the communities of justification—that is, the publics—within which all reasons (moral or otherwise) become effective.

6.4 The conditional validity of public reasons

To this point, I have been arguing that if people act on public reasons, they are likely to engage in collective action. But should they act on public reasons? That is another question. It may seem as if the answer is obviously “yes” since, given the basic structure of collective action problems, those courses of action that are justifiable on the basis of public reasons are likely to be better for everyone involved than actions urged by private or partial considerations.

But in fact, acting on public reasons is not necessarily the right choice because the validity of public reasons depend upon whether there is, in fact, a public. As I pointed out above, the idea of something being a reason—a claim with some binding or compelling force—only makes sense among people who take themselves to be accountable to each other in the first place. That means that in situations where the question of what to do turns on the validity of a public reason (say, a binding social norm or a shared goal), the rational choice cannot be specified without knowing whether, in fact, the people involved see themselves as accountable to each other. That question—are we relating to each other on a basis of mutual accountability or is this, in effect, a free-for-all?—becomes an essential part of the decision-making context. How people answer it determines whether they see public reasons in favour of cooperation as binding, and thus, ultimately, whether they are motivated to cooperate.
This is a surprising conclusion. It is surprising from at least two widely held points of view. First, from the perspective of rational choice models, rational choices should be describable in purely objective and individualist terms. An action situation, on this view, is fully described if we know the players involved, their goals and opportunities, and their knowledge of each other’s goals and opportunities.\textsuperscript{37} Within this framework, whether it is rational to cooperate when facing a collective action problem will depend on whether cooperation is part of an equilibrium strategy and, if there are multiple equilibria, whether the players have some way of knowing what strategy the other will play. The rationality of cooperation does not, in this framework, depend on the relational or intersubjective parameter of whether, and to what extent, the actors involved see themselves as accountable to each other. Indeed, it is difficult to see how such a parameter could even be conceivable within a rational choice model, given the instrumental conception of rationality such models imply. As I argued above, if rationality consists only in the choices of means to fixed ends, there is no way to express the possibility of actors being accountable to each other for the rightness of their ends or preferences.

But as I also noted above, the idea that the rationality of cooperation depends upon the relational context of interaction also conflicts with morally prescriptive approaches to collective action problems. A morally prescriptive approach would be one that sees cooperation in such situations as a rule of practical reason. On this view, when faced with a situation in which individuals must choose between cooperating for the common good and self-interested defection, cooperation is simply the right thing to do. From a morally prescriptive standpoint, practical rationality consists in acting in accordance with right reasons. It is the rightness of those reasons

\textsuperscript{37} Osborne and Rubinstein 1994, 11–13.
themselves that determines the rational choice and not, at least in any obvious way, the relationship of mutual accountability between the actors or the existence of a public.

By contrast, the conception of public reasons and collective action I have presented here recognizes that where public reasons are in play, the question of what to do depends on irreducibly intersubjective or relational realities. This represents a departure from all attempts to specify rational choices in objective and individualistic terms, regardless of whether those attempts refer to instrumental reasons or normative ones. The point is that from a purely individual point of view, rational action in the face of a collective action problem is fundamentally indeterminate. The question, “What should I do?” cannot be answered without reference to the question, “Are we (meaning me and the other people involved in this problem) relating to each other on a basis of mutual accountability?”

6.4.1 Understanding failures of collective action

One of the benefits of the framework I am defending here is that it helps us to recognize failures of collective action that are otherwise hard to understand. There are two common ways of explaining what goes wrong when people fail to cooperate in the face of collective action problems. The first is to see such failures as a tragic but inevitable product of individually rational choices. This is the view implied in game theoretical approaches to collective action problems. On this view, given the arrangement of incentives people face, defection is (if in a perverse way) the rational thing to do. Thus, people cannot be faulted for it, nor should we expect them to do anything else. The only possible response is to try to change those incentives, whether by introducing new sanctions or monitoring, or perhaps socializing people so that they have different preferences.
The second possibility is one that I highlighted at the end of Chapter 5. Here, tragic outcomes are the result not of *rationality* but *irrationality* in the form of “opportunistic defection.” Opportunism is a form of practical irrationality in that it involves a failure to choose the greater good over the lesser; an opportunistic actor will defect for the sake of securing gains now even though in doing so she looses the greater benefits of cooperation in the future. In cases of opportunism, the actor herself recognizes (if not in the moment of choice, then some time after) that the choice was not the right one: that in some sense, her desire for short-term gains got the better of her. In this way, opportunistic defection is on par with addiction, procrastination, and other forms of practical irrationality that can be recognized as such with reference to the actor’s own judgments.\(^\text{38}\)

But the framework I have established in this chapter brings into view a third possibility distinct from these. Some tragic outcomes are rooted not in failures of individual rationality or tragic circumstances, but rather in the failure of individuals to properly recognize and respond to the *relational* context of their decisions. These occur when people do not commit to public reasons because they do not realize that the relational prerequisites for those reasons are in place. In these cases, actors *individually* recognize a common good or cooperative plan and see it as a compelling reason for action; and yet they do not see themselves as a public, that is, they do not see themselves as being accountable to each other. Despite the fact that they each take the common good or cooperative rule to be a good reason (that is, their reasons overlap), they do not think of themselves as being accountable to each other in terms of that reason. As a result, the overlapping reason fails to be taken as a public reason, and thus fails to give rise to collective action.

\(^{38}\) Korsgaard 2008d; Heath and Anderson 2010.
6.5 Relationships and behavioural expectations

Readers might wonder whether the argument of the last section is quite right. Are people facing collective action problems really concerned about relationships of mutual accountability, or are they simply concerned, in a more straightforward way, with what each other will do? If cooperation in collective action problems is conditional simply on behavioural expectations and not relationships, then the relational conception of acting for a public reason would be an unnecessary theoretical addition.

At first glance, the suggestion seems plausible. The real problem actors face in collective action, and the reason they will often be reluctant to cooperate, is the threat that others might sucker them. What matters is what others do: it is the actions of others that will benefit or harm the decision-maker, and thus the decision-maker chooses with the likely actions of others in mind. This is certainly true if we approach collective action problems from a rational choice perspective. But it is also true for people who include social norms in their decisions. Social norms generally prescribe conditional cooperation: one should cooperate with other cooperators, but defect when facing defectors or cheaters. Whether people should cooperate in the face of collective action problems depends on what others are likely to do. Thus, behavioural expectations are doubtless an important part of reasoning in the face of collective action problems. We should not expect political opponents to cooperate unless they expect cooperation in return.

---

39 This, for example, is embodied in the “Tit-for-Tat” strategy discussed by Axelrod (1984). Norms of conditional cooperation generally do well in simulations of evolutionary dynamics (Bicchieri, Duffy, and Tolle 2004). Conditional cooperation is also supported by what behavioural game theorists call “strong reciprocity” (Falk and Fischbacher 2005).
Nevertheless, to reduce the interactive element of practical reasoning to behavioural expectations is too simple. People who cooperate expect others to cooperate too, but that expectation is itself formed through a conceptually prior assessment of the relationship they stand in to others. The word “expect” has both a predictive and normative sense. My argument is that people expect each other to cooperate in the predictive sense because they expect each other to cooperate in the normative sense. And developing that normative expectation—that is, adopting an attitude of mutual normative accountability toward the other—is where the action is.

The argument here runs parallel to the argument I made above in section 6.2 about the relationality of public reasons. What makes reasons public is not the fact that they happen to be subscribed to by more than one person. In fact, the mere overlap of people’s reasons is neither sufficient nor necessary for those reasons to be public. What makes reasons public, rather, is that they are addressed to each other as reasons, which implies a practice of mutual accountability between the people involved. While there is a connection between the publicness of reasons and their overlapping endorsement, the connection is indirect and contingent rather than necessary.

The same is true of public reasons and the actions of others. Predictive expectations that others will cooperate matter, but they are arrived at largely through the formation of normative expectations, which are founded on the assumption that others are part of an accountable public. The perception of accountability is conceptually and inferentially prior to the more specific behavioural expectations. And that priority can be demonstrated by looking at what happens when cooperative actions and relationships of accountability come apart. Such cases show that

40 As Neblo (forthcoming, 127) argues, a norm is “a pattern of action that is reasonable to expect of most people (in the predictive sense) largely because most people believe it to be a reasonable expectation (in the normative sense).”
correct predictive expectations about other’s actions are neither sufficient for, nor necessary to, the rationality of acting on public reasons.

The *insufficiency* of predictive expectations has to do with the problem posed by what Rawls calls a “mere modus vivendi.” A modus vivendi is, in effect, cooperation without normative accountability, cooperation that is merely a marriage of convenience. In such an arrangement, participants cooperate because doing so is, at least for the moment, advantageous, but they do not feel any *obligation* to each other to cooperate. The moment cooperation ceases to serve their interests, they will defect without scruple, because they have no commitment to the cooperative scheme beyond that derived purely from its benefit to themselves.42

If cooperation is a mere modus vivendi, then reasons prescribing cooperation are not public reasons. The danger of a modus vivendi is that a naive and well-meaning political actor would mistake this form of cooperation for one that involves genuine normative commitments between the players. When the winds shift and cooperation falls apart, the mistake can be a costly one.

The distinction between a mere modus vivendi and cooperation involving mutual accountability is implicitly reflected in everyday political practice. It is reflected in the tendency of wary political leaders to worry about what each others’ “real” intentions are even when they cooperate with each other.43 Such concerns reflect the fact that there is an important difference between an opponent who does what you expect only because it is useful to them and someone

---

41 Rawls 2005, 147.

42 The move away from a mere modus vivendi is also the key characteristic of a security community (Risse 1996; Adler and Barnett 1998). In a security community, peaceful relations and norms of non-violent conflict resolution have become so ingrained that the choice not to engage in aggression toward other members of the community is simply no longer a matter of instrumental calculation.

who does what you expect in part because they see themselves as being accountable to you in
terms of shared reasons. And that difference, in turn, makes a difference for how you should act.
If cooperation is only a modus vivendi, then it would be a mistake to take apparent obligations to
the other seriously as reasons to act, because the relational prerequisite for those reasons—
namely, a relationship of mutual accountability—does not in fact exist.

At the same time, other situations show that correct predictive expectations are not
necessary for it to be rational to act on public reasons. These situations are the converse of the
mere modus vivendi: rather than cooperation without accountability, you have accountability
without cooperation. This happens when political opponents disappoint each other’s behavioural
expectations, and yet have a good reason for doing so. Public rules (like all rules) nearly always
have implicit “unless” clauses attached to them: they commit you to doing the prescribed
action, unless a legitimate excuse can be made. Changed circumstances, inability to comply,
genuine misunderstanding or ambiguity in the meaning of the rules—all of these can, depending
on the circumstances, excuse deviations from the letter of the law. When they do so, it usually
makes sense for others to go on cooperating in accordance with public reasons, even though their
specific behavioural expectations were disappointed.

The interesting point is that it is not merely what the other does that matters; it is also
why he does it. Political opponents are concerned, in other words, to distinguish between dissent,
good-faith disagreement about what, in fact, they owe to each other, and deviance, the deliberate
disregard of obligations per se. Which it is matters for how other actors should respond. In the

45 This is reflected in the principle of international law, pacta sunt servanda rebus sic stantibus: treaties are to be observed, so long as conditions remain the same. Schroeder 1994.
46 Heath 2001a, 152.
case of dissent, the proper response will include argument and clarification about what the rules are, and perhaps an adjustment in action based on revised judgments about what public reasons imply. But if it is deviance, the relationship of mutual accountability that underpins the validity of public reasons is damaged or even destroyed. In such cases, the proper response will likely be coercive, not argumentative.

Thus, while it might seem, at first glance, that the validity of public reasons is straightforwardly conditional on expectations about what others will do, on closer examination, something else is going on: acting for public reasons ultimately depends on the relational question of whether actors see themselves as accountable to each other. That is not to say, of course, that people facing collective action problems are not concerned with what others will do. They are. But those behavioural assessments themselves are conditioned by judgments about others’ attitudes of mutual obligation and commitment or lack thereof, and cannot be reduced to predictive expectations about what others will do.

6.5.1 Experimental support

There is, in fact, some support for this conclusion in laboratory evidence based on how people play one-shot Prisoners’ Dilemmas. Eldar Shafir and Amos Tversky describe an experiment in which they asked subjects to play a standard one-shot Prisoner’s Dilemma.\(^\text{47}\) The twist was that, unlike in a normal Prisoners’ Dilemma, some subjects were told what action their opponents had taken before they made their own choices. When the subjects knew the other had defected, they cooperated 3% of the time; when they knew the other had cooperated, they cooperated 16% of

the time. The surprise is that when the players did not know each other’s move (i.e. a normal Prisoners’ Dilemma), the cooperation rate jumped to 37%.

How can we make sense of this result? Within a standard game theoretical framework, it does not make sense: the cooperation rate should simply have been zero in all three trials, since defection is the dominant strategy in a one-shot Prisoners’ Dilemma.

More germane to my argument here, though, is the fact that these results also do not fit with the hypothesis that some of the subjects were basing their cooperation on a *predictive* expectation that others would cooperate too. In particular, that hypothesis does not fit with the fact that cooperation was much higher when the subjects didn’t know what the other had done than when they *knew* the other had cooperated. If people were following a norm of conditional cooperation, and deciding based on a prediction about the others’ behaviour, the cooperation rate should have been highest when they knew with certainty that the other had cooperated.\(^{48}\)

Then what can account for the strange jump in cooperation when people did not know the other’s action? One thing that could explain these results is if a large portion of the subjects were conditioning their actions not on direct predictions about the other’s actions, but rather, as I have argued in this chapter, on whether they perceive themselves to form a public. On this view, rather than asking themselves “will the other cooperate or defect?” they ask “are we acting on public reasons or only on private reasons?”—or, simply, “are we following shared rules or is this a free-for-all?” This changes the significance of whether or not the subjects know what action the other has taken when they make their own choices. If, when choosing, the players already know what the other has done, they think of themselves as acting individually on the basis of private

\(^{48}\) On this view, the subjects would be violating the “sure thing” principle (Chapman 2003, 1184; the original source is Savage 1954.)
reasons, and are thus more likely to follow the standard strategic argument for defection. But if they do not know what the other has done, it is easier for them to think of themselves as acting together with the other. At the moment of choosing, the other’s action is still, in a sense, undetermined or up in the air, and thus it is easier to think in terms of shared goals or rules. In other words, it makes it easier to think of themselves as a public. As a result, they are more likely to cooperate.

6.6 Conclusion

In this chapter, I have argued that relationships of mutual accountability are the key to cooperation in the face of collective action problems. Where people see themselves as being accountable to each other for the rightness of the reasons they act on, new ways of acting together become possible. Collective action is about the sharing of practical reasons or intentions: not the incidental sharing implied in individually overlapping strategies, but a sort of sharing predicated on mutual normative accountability. When that accountability is in place, public reasons become valid and reliable ways of coordinating action. Thus, the challenge of collective action is ultimately the challenge of constituting a public.

Establishing this connection between public reasons and collective action is a key step toward developing a deliberative theory of cooperation in anarchy. Not surprisingly, there are good reasons to think—as I will argue in the coming chapters—that deliberation plays a crucial role in establishing and thickening relationships of mutual accountability between political opponents. Deliberation is, in effect, the practice of mutual accountability—a practice that consists in the give and take of reasons that are essentially public. That is the argument I will take up in the following chapter.
In the last two chapters, I have established a key connection between public reasons and cooperation. Cooperative rules generally take the form of public reasons—practical reasons in terms of which actors take themselves to be accountable to each other. Where people come to treat their practical reasons as public, they become responsive to each other in a way that goes beyond the merely strategic responsiveness modelled in rational choice approaches. The introduction of public reasons amounts to a new way of interacting—one in which each person’s actions are structured by mutual accountability and shared standards of justification right down to the level of preferences themselves. Action structured by public reasons and mutual accountability differs fundamentally from purely instrumental, strategic interaction.

How should we understand deliberation in light of these findings? It is possible to conceive of deliberation in purely instrumental and strategic terms. Deliberation may simply be another form of strategic struggle, in which participants attempt to sway events by choosing what to say, and what not to say, with a view to altering each other’s strategic choices. This certainly seems like a plausible view. There is often much at stake in deliberation, much to be won and lost, and it is only natural to think that people will choose their words so as to bring about the results they want. Deliberation is about the give and take of reasons, but those reasons are not neutral with respect to the decision at hand. Rather, reasons drive toward conclusions; they shift
the argumentative landscape to make some conclusions seem more appropriate than others.\textsuperscript{1}
Deliberation can be construed as simply another form of strategic struggle: war by other means.\textsuperscript{2}

As plausible as this view of deliberation might seem, however, it is fundamentally misguided. Deliberation cannot be simply another form of purely strategic interaction because, as I will argue below, the reasons exchanged in deliberation are necessarily public reasons. Mutual accountability is, in fact, at the heart of deliberation. To remove that accountability—which is what happens when deliberation is construed as purely strategic interaction—is to lose what is most distinctive about deliberation as a form of communication: its ability to communicate not only information but also inferences. Deliberation is essentially a practice of holding one another accountable in terms of shared norms and standards of justification.

But while deliberation is predicated upon mutual accountability, it is also clear that not all communication is. Words can be used in ways that do not depend upon accountability between speakers—and indeed that actively undermine that accountability. That is precisely what happens when people treat reasons, their own and others’, as private rather than public. In that case, communication is restricted to non-inferential modes of signalling and coordination. Not all language use is purely instrumental, but some certainly is.

How, then, should we understand the contingency of normatively structured and purely instrumental language use? I will argue that the answer here parallels the findings of the previous chapter. The conditions for speaking through public reasons are the same as for acting on public reasons: participants must perceive themselves as standing in relationships of mutual accountability with others. In other words, deliberation depends upon the existence of a public.

\textsuperscript{1} Cf. Calvert 2013.
\textsuperscript{2} Krebs and Jackson 2007; Schimmelfennig 2001; Tully 1988.
But a public is a contingent social achievement. It exists only where people have escaped what Kant called a “state of war”: a situation in which conflicting groups have not yet accorded each other the basic recognition of mutual responsibility that makes reasoning with each other a possible response to conflict.

Establishing this inherent connection between deliberation and mutual accountability is an important step toward explaining deliberative influence in anarchy. As I will argue in the next chapter, much of the work deliberation does in relatively anarchical circumstances lies in building relationships of mutual accountability—relationships that make possible cooperation on the basis of shared rules and commitments.

The chapter will be structured as follows. I begin, in section 7.1, by summarizing the pragmatic philosophy of language of Robert Brandom and Jürgen Habermas, which provides the conceptual underpinning of deliberative theory. This involves two main ideas. The first is inferential semantics: the idea that the meaning of words (or, properly, utterances) is a property of the role they play in reasoning or inference. The second idea is that reasoning is a normatively regulated social practice. The conceptual content of a statement is constituted by the way speakers and listeners attribute commitments or entitlements to each other to make that statement. Mutual accountability is thus essential to the practice of reasoning together.

In section 7.2, I argue that this inferentialist or reason-giving aspect of communication is only conceivable if reasons are understood to be public. Rational choice models of communication, such as cheap talk and costly signalling, represent reasons as private. As a result, they fail to capture what is most distinctive about deliberation as a form of communication: its ability to communicate inferences as well as information.

3 Kant 1991b, 165.
Section 7.3 argues that this limitation applies to all models of communication that are not based on public reasons and mutual accountability. As long as we think of people’s reasons as private, individual reasoning remain fundamentally disconnected from collecting reasoning. The resulting problems of judgment aggregation and incentive incompatibility can only be resolved if reasons are recognized as public rather than private.

Finally, section 7.4 qualifies and refines the argument of the first three sections. It concerns the difficulty of squaring the pragmatic philosophers’ argument that language use is inherently normatively structured with the apparent ubiquity of anomic or strategic uses of language, especially in politics. The answer, I argue, is to keep the levels of analysis clear. Brandom and Habermas are right to argue that language in general (and, indeed, rational thought itself) is inherently normatively structured. In this sense, it is not possible for linguistic practice to be wholly instrumental at the global level. However, it is possible for language to be used in an anomic and strategic way within particular contexts of interaction. Or, to put it another way, not all reasons are public, though perhaps all languages are. Sometimes people fail to treat their reasons and the reason of others as public, and when they do, their language use will depart from the conceptual ideal of deliberation that is articulated in philosophical pragmatism. Understanding the contingency of normative commitment is essential to politicizing the pragmatic philosophy of language.

7.1 Communication as deontic scorekeeping

Deliberative democratic theory is rooted in a particular way of understanding how communication works. At heart, deliberation must be understood as a social or relational practice—that is, as a normatively regulated, socially constructed way of doing things.
Communication is also, of course, about things. Meaningful statements have propositional content: content, that is, that refers to something else, something beyond the words themselves. But this “aboutness” or referential aspect of speech is not all that is going on when people communicate with each other. Communication also has a pragmatic and relational structure. In speaking—for example, making a statement or asking a question—speakers and listeners establish, assume, or modify social entitlements and statuses. It is ultimately this relational structure of communication, the significance uses of words have for a social structure of interlocking commitments and entitlements, that allows this theory of communication to give rise to the social and political theory that deliberative democrats work with.

In this section, I will explain this idea of communication as a social practice. The argument can be made in two parts: first, the argument that meaning is a function of inference or reasoning; and second, that inference emerges out of the social attribution of deontic or normative statuses.

7.1.1 Meaning is determined by structures of inference

The first part of the pragmatic conception of language is inferential semantics. That is the idea that the meaning of words (or, properly, utterances) is determined by the role they play in valid inferences. “Inference” here means reasoning: the correct movement from premises to conclusions. Any given utterance has inferential connections running in two directions: it can stand as both a conclusion in need of justification, and as a premise that justifies other

---

4 Habermas 1998; Brandom 1994.
conclusions. To know what an utterance means is to have a grasp of both these sets of inferential connections: inferential conditions and inferential consequences.  

For example, you might look out the window and say to me, “it’s raining.” On this inferential understanding of semantics, I have understood the meaning of your statement if I have grasped the role that statement might play in chains of reasoning or inference. In one direction, those chains reach backward to give the inferential conditions, the premises, that would justify the conclusion, “it is raining.” You might, for example, have seen raindrops splashing in puddles and people putting up umbrellas. There are many premises or observations that might justify your inference that it is raining. In the other direction, those chains of inference reach forward to give the inferential consequences, the conclusions, we would be justified in drawing from the premise, “it is raining.” If I know it is raining, for example, I might infer that I would get wet if I go outside or, perhaps, that I should have closed the window before I left the house this morning. In sum, when you say, “it is raining,” I have grasped your meaning if I know what your statement would follow from, and what would follow from your statement. The meaning of a statement is determined by the role it can play in chains of reasoning.

The main alternative to an inferential account of meaning is a referential one. On this view, to know what a statement means is to know what it refers or points to. Thus, if you say to me, “it is raining,” I have grasped your meaning if your statement conjures in my mind the particular state of affairs you were referring to, namely, to the fact that it is raining. The meaning of the statement consists of this pointing: the sense in which your statement is about a state of affairs in which drops of water are falling from the sky.

____________________________

5 Brandom 2000, chap. 2.
The choice between an inferential and referential semantics should not be taken as a mutually exclusive one. Indeed, both Habermas and Brandom, both proponents of inferential semantics, also see meaning as involving representation. Habermas distinguishes different kinds of validity claims according to which “world” they represent: truth claims refer to the objective world of facts; rightness claims refer to the social world of norms; and truthfulness claims refer to the subjective world of inner experience. Brandom argues that representation is a “dimension” of meaning, though one that can ultimately be explained in terms of inference.

Rather, the choice between inference and representation is, as Brandom argues, a choice about explanatory priority. On the one hand, we can start with the concept of inference and derive from that the content of what is said. On the other hand, we can start with the concept of true representation and derive from that an understanding of inference. The inferentialism of Habermas, Brandom, and other pragmatists is a choice of the first explanatory strategy over the second. The choice, however, does not have to be a reductive one. A successful inferential or pragmatic account of meaning would be one that explains, not explains away, the representational aspect of meaning.

To justify this explanatory priority in philosophical terms is beyond the scope of this dissertation. For my purposes (and, I suspect, political theory generally) it is enough to recognize the connection between meaning and inference: to recognize that knowing what something

6 Habermas 1984, 1:51.
7 Brandom 2000, chap. 5.
8 Brandom 2000, 10–12.
9 Brandom 2000, 28.
means involves knowing what it follows from, and what follows from it. For that, all that is needed is a semantics that does not reduce meaning purely to its representational dimension.

### 7.1.2 Inference as a social practice

The second part of the argument has to do with the social nature of the inferential proprieties that give words their meanings. To see this, begin with the claim that was central to the conception of practical rationality presented in Chapter 5: the claim that reasoning is essentially a normative phenomenon. Reasons, that is, are “oughts”: if you have a reason to do or believe something, then you *ought* to do or believe that thing (barring, of course, reasons to the contrary). Valid inferences are ones that I am entitled or obligated to make; invalid inferences are ones that I am not entitled to make. (And again, this basic normativity or prescriptiveness is a feature of all reasons, not only those that are normative or moral in content.)

Thus, if you tell me it is raining, and I believe you, it follows that there are other statements that I ought to see myself as committed to, as a matter of right reasoning. I ought, for example, to expect to get wet when I step outside. If I do not make that inference, then something is amiss. I have not understood the meaning of your words, and that means that I am not entirely as I should be as a competent user of the English language. On the other hand, there are other inferences I might make that I should not. Thus, if you tell me it is raining and I conclude, “well, it must be a Tuesday,” I would be making an inference to which I am not entitled—a conclusion that I ought not to draw.

---

10 Chapman 2005, 106.
What I am referring to here is simply the basic normativity inherent in all reasoning. But the pragmatic philosophy I am discussing here implies something beyond this. It implies that the source of this normativity lies in social practices of judgment.

This can be represented as a move from Kant to Hegel.\textsuperscript{11} Kant made normative bindingness central to his conception of rationality. Kant noticed that right reasoning has a sense of necessity about it, something that is especially clear when we feel ourselves “forced” by evidence to believe things that we’d rather not believe.\textsuperscript{12} Rationality, on this view, is about acting or believing in accordance with those reasons that we take to be right.

But what is the source of this normativity, this binding quality good reasons seem to have? This, arguably, is where Hegel comes in.\textsuperscript{13} Hegel’s insight was to see normativity as a social achievement. The bindingness of rational judgments that Kant noticed is seen by Hegel to be a product of social practices of attributing normative statuses to each other.

This helps to explain something that is otherwise mysterious on a Kantian view. Kant locates the bindingness of correct inferences in the actor’s endorsement of them as binding. This is Kant’s idea of autonomy. The only type of laws or imperatives that are consistent with the freedom of the thinking, choosing being are laws that she gives herself. The problem, though, is this: if correct inferences are binding only by virtue of the individual’s taking them to be binding, are they really binding after all? How can I bind myself? One plausible answer is that I get help

\textsuperscript{11} Brandom 2000, 34.

\textsuperscript{12} Kant 1997, 27; Neblo forthcoming, 139.

\textsuperscript{13} I say \textit{arguably} because Kant, too, appeared to be moving toward a more social or intersubjective conception of judgment in his political writings and in his third Critique. Kant 2000; Arendt 1982.
from others. The “bindingness” of my linguistically-incurred commitments comes both from my undertaking them (the Kantian part) and from other people’s attributing them to me.¹⁴

Brandom sums this up nicely in his notion of language as a game in which participants engage in “deontic scorekeeping.”¹⁵ Each statement I make changes the “score” of what other commitments I now have by virtue of having made that statement. The key point is that this scorekeeping is a social practice. It is not just me attributing a score to myself (as it would be if language, and reasoning, were a wholly private activity); others attribute a score to me. It is precisely the social aspect of normative scorekeeping that gives words determinate conceptual content, content that is not just whatever I want it to be. Thus, as Brandom puts it,

I get to decide which piece in the game I will play—say, the one labelled “That metal is molybdenum,” or “I promise to drive you to the airport tomorrow morning,”—but I do not then get to decide what I have committed myself to thereby, what further moves are appropriate or obligatory for one who has played that piece.¹⁶

The essential point of this philosophy of language, then, is that communication involves a continual exchange of normative statuses and commitments between speakers. Moreover, these normative commitments are not optional or incidental to communication—prescriptions we add after the fact.¹⁷ Rather, mutual commitments are the vehicle of communication itself.

¹⁴ Brandom 1999, 172.
¹⁵ Brandom 2000, 81.
¹⁶ Brandom 1999, 172.
¹⁷ Warren 2012a, 16.
What are the implications of this pragmatic philosophy of language for understanding politics? The most important implication is that the reasons exchanged in deliberation must be understood as *public* reasons. Pragmatism reveals communication to be an essentially normative social practice. The transmission of meaning takes place through the way interlocutors attribute responsibilities (that is, commitments and entitlements) to themselves and others. Mutual accountability thus lies at the heart of communication. It is precisely the publicity of reasons—the way they are addressed to each other as claims about what speakers are responsible to each other for—that allows them to carry intersubjective meaning.

Pragmatism thus provides a crucial link between the activity of deliberation, on the one hand, and the constitution of a public, on the other. There is, in other words, an essential conceptual connection between deliberation and the relationships of mutual commitment that constitute a group of actors as a public. Furthermore, as I argued in the last chapter, the constitution of a public is the key to collective action. When people act for public reasons, the publicity of those reasons means that their actions can be structured by shared accountability to collective goals and norms, resulting in actions that cannot be achieved through purely private reasons.

To be clear, the conceptual connections between deliberation, publicity (i.e. mutual accountability), and collective action have not yet been spun out into causal and sequential claims. That is to say, we know that deliberation and mutual accountability are rationally or conceptually implicated in each other, but that does not yet say which (if any) comes first, or how one might lead to the other. I will take up that question in Chapters 8 and 9. But before causal hypotheses involving deliberation can be advanced, we must first get clear on what,
conceptually, is implied in deliberation and why we might expect it to be implicated in relationships of mutual accountability or collective will-formation.

7.2 Deliberation and rational choice

The claim that deliberation must be an exchange of public reasons is supported by looking at what happens when communication is modelled as the exchange of private reasons. This, in effect, is what happens in game theoretical models of deliberation. As I argued in Chapter 6, game theory necessarily models the practical reasons of actors as being private. The result, when applied to deliberation, is to assimilate inference to information, and communication to signalling. In such models, what is most distinctive about deliberation as a form of communication—namely, reason-giving—is lost.

7.2.1 Cheap talk and costly signalling

To see this, consider the two best-known ways of modelling communication in game theory: cheap talk and costly signalling. Cheap talk occurs in the context of coordination problems, situations in which there are multiple equilibria available and the actors are thus uncertain about which strategy each other will play. Meeting in a cafe is a classic example of a coordination problem. \(^18\) Two friends want to meet for coffee. Each has a choice of going to one of two cafes. Each is indifferent as to which cafe they go to, as long as they end up at the same one. In this situation, there is clearly a role for meaningful communication. One friend can simply tell the other which cafe she is going to, or they can talk about it and come to an agreement; then going to

\(^{18}\) Heath 2001a, chap. 2; Chapman 2006.
the agreed-upon cafe is clearly the rational choice for each to make. Communication serves as a “focal point”; it makes one equilibrium salient and allows the actors to coordinate on it.\textsuperscript{19}

Furthermore, the idea of using words to coordinate in the face of multiple equilibria has been suggested as a model not only of discrete instances of communication, but as a general model for the development of language itself.\textsuperscript{20} On this view, language is an elaborate convention, a set of symbols, that allows actors to coordinate on mutually beneficial strategic equilibria. Like the rule of driving on the right, linguistic symbols, once established, are self-enforcing equilibria.

The problem with this idea, however, is that such “cheap talk” is highly vulnerable to payoff asymmetries. Cheap talk works best when the people involved are indifferent as to which of the multiple equilibria is played, like friends who don’t care which cafe they go to as long as they go to the same one. But when it comes to situations of conflict, such indifference is unlikely. Different equilibria benefit some players more than others. And where such a payoff asymmetry is in place, the ability of cheap talk to successfully convey information quickly breaks down.

For example, government armies and rebel groups who want to avoid bloodshed (at least for the moment) sometimes attempt to coordinate the locations of their patrols. Governments and rebels sometimes concede de facto zones of control or spheres of influence to each other in this way.\textsuperscript{21} If the two sides were indifferent as to where the line of control fell, one side could simply tell the other where it will patrol. But where exactly those zones of control fall is not likely a

\textsuperscript{19} Schelling 1960.

\textsuperscript{20} Heath 2001a, chap. 2.

\textsuperscript{21} Staniland 2012.
matter of indifference to both sides. Suppose that there is a particular valley close to the line of control that contains valuable minerals that could be mined by the rebels. Now, when the rebel group announces that it will patrol the valley, the government cannot infer from the statement what the rebels’ intentions really are, because the rebels have an incentive to say that it will patrol the valley regardless of whether that is true. Thus, the group’s statements lose their ability to convey information.22

These game-theoretical results have been corroborated with experiments. In one study, subjects played a coordination game in which they had to choose which of two Chicago locations to meet at: the well-known Sears Tower or the lesser-known AT&T building.23 They were each paid $100 if they met. Predictably, 90% choose the Sears Tower and the players received the payoff in 82% of the games. But with the introduction of a very small payoff asymmetry—one player received $100 dollars if they met at the Sears Tower and $101 if they met at the AT&T building—the coordination rate dropped to only 50%. This is an astounding difference given that the asymmetry of $1 is so little in comparison to the $100 both lose by failing to coordinate.

Costly signalling, the second common mechanism for communication in game theory, addresses this problem. If there are costs attached to speaking, those costs—and the fact that an opponent is willing to send the signal in spite of them—may allow others to make reliable inferences about the speaker’s own incentives, and thus about what strategy the speaker will play.

---

22 Heath 2001a, chap. 2; Landa and Meirowitz 2009, 436; Morrow 1999.

23 V. P. Crawford, Gneezy, and Rottenstreich 2008.
Costs can take many forms. “Audience costs” are one well-known example. In crisis diplomacy, for example, where two states are making competing demands and there is a threat of military escalation, declarations of resolve—“if you don’t make that concession, I’ll take military action”—would normally be undermined by the problem of cheap talk. But if claiming to be resolved when you aren’t has costs attached to it, the situation is different. For example, a democratic leader who pledges to stand up to an enemy and then later backs down might fare worse at the ballot box than one who simply backs down. This cost then gives the opponent new information: since a leader who bluffs is likely to be punished, a declaration of resolve is evidence that the leader making it is not bluffing—that is, he really is willing to fight. Likewise, a costly signal might also be used to dampen conflict by signalling a preference for negotiation and peace. Here, taking a risk for peace provides the signal. A risky concession can be evidence that the actor is genuinely committed to peace if an actor who was less interested in peace would not have borne that risk.

7.2.2 Inference versus information

What I want to point out about both cheap talk and costly signalling models of communication is that what is communicated in both these models is information but not inferences. By information I mean discrete facts. In costly signalling, one state might learn something about another state’s preferences, and thus about what actions it will likely take; in cheap talk, players

may learn what actions each other intends to take. Inferences, by contrast, are not just discrete facts, but rather *claims* that link facts together as a piece of reasoning.

This marks a fundamental difference between cheap talk and costly signalling, on the one hand, and deliberation, on the other. Communicating inferences, and not just revealing information, is the essence of deliberation. In all the quintessential deliberative “moves,” the validity of inferences themselves is the subject of communication, as opposed to discrete facts or information. Offering a reason for why an opponent should accept a claim, demanding that an opponent justify a claim she has made, and accepting or rejecting the reasons an opponent offers: each of these moves involves either asserting or questioning the inferential connections between premises and conclusions. To offer a reason for a claim is to treat that claim as a conclusion for which you are making clear the premises. To demand justification of another’s claim is to treat that claim as a conclusion standing in need of premises. To accept a reason is to endorse the inference it contains as good or valid, while to reject a claim is to reject the validity of that inference. Inference is the engine of deliberation.

To be sure, inference is also necessary for successful cheap talk and costly signalling. These modes of communication require the hearer to make inferences from the fact that sending the signal was in the speaker’s interest. When one state signals resolve by undertaking a costly military deployment in the middle of a crisis, the signal works because the receiver can infer the sender’s motive from the fact that it is willing to bear that cost. The sender in effect transmits information about himself through inferences the receiver makes. So how are the inferences at play in deliberation different?

The answer is that in deliberation, *it is the validity of the inference itself that is communicated*, and not (or not only) information to be fed into previously existing inferential
structures. These represent two different mechanisms of communication, and two different ways in which our beliefs and commitments can be changed. On the one hand, we can update our beliefs by plugging new or different information into already-established structures of inference. On the other hand, our structures of inference themselves can change, as we realize that the information we have in fact entails conclusions that we hadn’t seen before, or that conclusions that we thought followed from that information in fact do not.

In cheap talk and costly signalling, the updating that takes place belongs only to the first type. Here, the structure of inference that both players bring to the interaction is fully worked out, fully transparent to themselves and each other. It is axiomatic in game theory that both players are “fully rational” in the sense that the way they process information is completely specified. That is necessary in order for players to be able to recognize an equilibrium strategy as an equilibrium strategy. Each must be able to mirror the reasoning of the other, and thus what reasoning consists in—that is, the structures of valid inference possible—must be fully worked out and shared between them. Thus, the signalling that takes place can be a matter only of revealing information: giving the receiver something new to plug into the established inferential structures. Or, to make the same point another way, the act of signalling reduces the receiver’s uncertainty—if there weren’t uncertainty, there would be nothing to signal—but uncertainty that comes from the player’s lack of information and not from what inferences to draw from that information if she had it.

27 The most common specification for making inferences is Bayes’ Law, but many more complex learning models are possible (see, e.g., H. P. Young 2004). Whatever the model, though, how the players make inferences—that is, what counts as a good inference—must be fully specified.

28 Heath 2001a, 269.
Deliberation, however, involves the second type of updating as well. To say that people are reasoning with each other is to imply not only that they are revealing or pooling information, but also that they are clarifying and modifying inferences. Deliberation is needed precisely where inferential structures are not fully worked out or fully transparent. Deliberation is about weighing different arguments against each other, about making our inferences explicit, openly confronting and questioning them, accepting or rejecting them. Deliberation involves a process of thought and decision in which what follows from what (and what doesn’t) is not fully clear to begin with. The sort of Socratic inquiry that represents one ideal of deliberation—the careful process of making arguments clear, of following inferences where they lead and then questioning the results—would not be necessary for people who were “fully rational” in the game theoretical sense, if, for example, their reasoning could be specified in terms of Bayesian updating. Arguments that wear their inferences on the surface, and where the inference itself, rather than just information, is being communicated, are typical of deliberation. “You believe A, B, and C. Don’t you see that you have to believe D, then, too?” Or, “You claim that A is right because of B, but you haven’t considered the effect of C.” In arguments like these, convincing an opponent is often not only a matter of revealing information, but of showing the opponent what is implied in the information she has. To conceive of deliberation purely in terms of the pooling and aggregation of information is thus to overlook the most distinctive work deliberation does.

29 Dietrich and List 2013.
30 Brandom 1994.
31 Brandom 2000, 72.
7.3 Deliberation and public reasons

Is this just a problem for cheap talk and costly signalling, the most common ways of modelling communication in game theoretical terms? Or is the problem a more fundamental one? I suggest that it is the latter. The fundamental limitation of rational choice treatments of communication is that they conceive of reasons as private, and that is why they are limited to information-based communication or signalling. For deliberation to be based on reasons, those reason must be public—which is to say that they must be addressed to each other as reasons, as claims or commitments that interlocutors are responsible for to each other.

Why might genuinely reason-based deliberation be incompatible with participants having only private reasons? The answer can be seen in the nature of the outcomes reason-based deliberation should secure. Where people deliberating a practical question succeed in persuading each other on the basis of reasons, the result will be the joint acceptance of shared reasons as shared. That is what it means to influence each other through deliberation or persuasion, as opposed, say, to a threat or deceit. The vehicle of deliberation is a successful claim about what is correct, what is owed to each other.

But such an outcome—a shared reason endorsed as a shared reason—is difficult to understand if reasons are exclusively private. Private reasons lack the accountability that is required for validity claims to get a grip on them. As I argued in the previous chapter, while private reasons might happen to shared in the sense that more than one person subscribes to them, they are not endorsed as shared. Their oughtness or conviction is a private matter, a matter of what the actor owes to herself or her own constituency, but not to others.

As a result, where reasons are private, collective outcomes remain disconnected from individual reasons. I use the word “outcome” intentionally, because that is part of the problem.
Individuals see the results of interaction as an outcome, but not as a rationally motivating will or intention. The aggregation of private reasons—whether through voting, signalling, or other strategic interactions—does not necessarily result in an outcome that is itself rational from the individual’s point of view. What is lost, in other words, is the possibility of collective will-formation: the generation of shared intentions based on reasons that individuals can recognize as simultaneously shared and their own.

In fact, we can use the conception of practical reason presented in Chapter 5 to be more specific about the problem of aggregating private reasons into collective outcomes. As I argued in Chapter 5, a practical reason has two parts: a motivation and the actor’s endorsement of that motivation as good. The disconnection between private reasons and collective outcomes manifests itself in both these dimensions of practical rationality. These are, in effect, two things that happen when individuals do not see collective reasons or judgments as their own. On the one hand, individuals will not endorse such outcomes as right; they will not carry a sense of conviction. On the other hand, individuals will often fail to be motivated to uphold collective decisions or reasons. As a result, deliberation will be stymied by incentive problems. I will explore each of these problems in turn.

### 7.3.1 The problem of judgment aggregation

First, consider the endorsement or judgment aspect of practical reasons. Here, the disconnection between private reasons and collective judgments takes the form of the “discursive dilemma.” The discursive dilemma arises when individual judgments that are implicitly understood as

---

private reasons are aggregated into collective judgments or decisions. The dilemma arises from the fact that reasoning that is correct according to collective reasons or standards of inference can result in outcomes that individuals do not find convincing and do not rationally endorse. The dilemma, then, is that we appear to be forced to choose between individual and collective rationality.

To see how it works, consider the “doctrinal paradox” identified by Kornhauser and Sager.\(^{33}\) A three-member court is deciding a breach-of-contract case. According to the law, the court’s decision should be governed by a particular inference or chain of reasoning: the defendant should be found liable if, and only if, (a) the defendant had a contractual obligation not to do (or not to omit doing) a particular action and (b) the defendant took (or omitted to take) the action in question. Suppose that the judges disagree on these various premises, as well as on the verdict. Judge 1 thinks the defendant had the obligation and did the offending action, and is therefore liable for damages. Judge 2 thinks the defendant had the obligation but did not do the offending action, and therefore is not liable. Judge 3 thinks the defendant did the action in question, but had no contractual obligation to the contrary, and therefore is not liable (Table 7.1).

\(^{33}\) Kornhauser and Sager 1986; Kornhauser and Sager 1993; see also Chapman 2003.
Table 7.1. The “doctrinal paradox”

<table>
<thead>
<tr>
<th></th>
<th>Obligation?</th>
<th>Action?</th>
<th>Liable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Judge 2</td>
<td>True</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>Judge 3</td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Majority</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>


How should the court decide the case? One way to do it would be to have the judges vote on the various propositions involved. According to proposition-wise, majoritarian aggregation, the judgment of the court would be that the defendant had the relevant obligation and took the offending action, but is not liable for a breach of contract.34 (See the last row of Table 7.1.) The problem, of course, is that this combination of propositions is inconsistent according to the legal definition of breach of contract. If a court were to hand down such a judgment, it would be clearly failing to do its job (and the plaintiff would have a right to be upset). In particular, it would be making a judgment that is not rational according to the standards of correct legal judgment. The court as a whole would not be able to give a proper reason for its finding of “not liable.”

So what is the alternative? The alternative is to allow the inferences enshrined in legal doctrine to determine the court’s conclusions. In other words, deliberation based on public

---

34 List and Pettit 2011, 43–47.
reasons—in particular, legal reasoning—should determine the outcome. In order to do that, the court would start by voting on the propositions that are premises—namely, Did the defendant have the relevant obligation? and Did he do the action in question?—and, finding that both are true, it would conclude, according to their shared standards of legal reasoning, that the defendant was indeed liable.

But this method of aggregation, too, is not without its problems. It does result in a decision that is collectively rational. Unlike in the first scenario, the court can now give an acceptable public reason for its verdict. But it appears to have achieved this collective rationality at the expense of individual rationality. While Judge 1 can wholeheartedly endorse the court’s decision, Judges 2 and 3 cannot. Judges 2 and 3 both see the decision as a mistaken one (though for different reasons). And they might understandably ask why the court should endorse a verdict that a majority of its members see as wrong. For Judges 2 and 3 to return a verdict that makes sense according to collective standards of reasoning, they have subsume or sacrifice their own rational judgments. They are being asked to buy into an outcome that they cannot, rationally, endorse. The collectively rational outcome remains disconnected from the judgments of those individuals.

7.3.2 The problem of incentives

Now consider the same problem from the point of view of the motivational efficacy of the shared reasons given in deliberation. Not only will a majority of judges fail to be convinced by the shared reasoning; they may also fail to be motivated by it.
Consider what happens when the judges are deliberating about the premises. Judges 2 and 3 might be tempted to try to steer the court away from the conclusion of “liable,” and they might do so by misrepresenting their own judgments to the others. Judge 2, for example, who is concerned that the defendant will falsely be found liable, might claim, insincerely, that she does not think the defendant had the relevant obligation. By doing so, she would shift the majority opinion on the question of obligation from “true” to “false.” Then the court, following the legal doctrine, would have to find the defendant not liable, which is Judge 2’s preferred outcome. (A similar incentive for strategic misrepresentation exists for Judge 3.) The sort of truthful communication upon which deliberation depends may, in such cases, be “incentive incompatible.” And assuming that all the judges are aware of these incentive problems, they will be unable to take each other’s statements at face value. As a result, their deliberations may break down.

This example illustrates the broader point that in order for deliberation to work, participants have to be able to trust each other to abide by shared rules of inference, as well as rules of sincerity and truth-telling. In many ways, this conclusion mirrors the problem with cheap talk: while “cheap talk” stands for the argument that actors’ ability to convey information can be undermined by incentive structures, a similar problem can affect inference.

That conclusion is not surprising given the pragmatic philosophy of meaning I outlined above. If inferential proprieties are indeed the vehicle of meaningful communication, then meaningful communication depends upon people using words in ways that respect those inferential proprieties. Communication depends upon respecting inferential rules, on drawing

36 Landa and Meirowitz 2009.
only conclusions that you are entitled to draw, and accepting the conclusions that are entailed in the statements you make. If people stop treating public reasons as binding—if they abide by them only to the extent that it is convenient for them to do so—the very meaningfulness of the statements they make to each other also breaks down. Rational persuasion only works when people respect the rational force, so to speak, of a valid inference. Take, for example, the straightforward inference \( if \ p \ then \ q, \ p \), therefore \( q \). The possibility of deliberation requires that when presented with this argument, a person who accepts the premises accepts the conclusion. If she does not—if she says, “yes, I’ve admitted \( if \ p \ then \ q, \ and \ p \), but I won’t accept \( q \)”—then the argument is going nowhere.

Of course, it is unrealistic that a person would flatly deny a basic principle of logic like modus ponens. But when it comes to the actual material inferences that political opponents make (as opposed to abstract logical principles), what constitutes a valid inference is often not clear cut, and it becomes quite possible to disagree about what follows from what. The ongoing debate over gun control in the United States illustrates the point. A proponent of gun control might argue thus: massacres are nearly always carried out using military-style rifles; if those rifles were harder to get hold of, massacres would be less common; therefore we should restrict or ban those rifles. But an opponent of gun control will not likely find this inference convincing. He may dispute whether the availability of assault rifles really makes massacres more common. He might argue that people who are determined to get their hands on such guns will find a way, and so bans and restrictions are largely pointless or symbolic. He might even draw the opposite conclusion: since deranged gunmen are out there and are capable of getting their hands on such
guns, it is important that upright citizens also be armed, so that when a massacre starts, people nearby might have a chance of stopping the gunman before even more people are killed.  

In actual conflicts like these, where arguments take the form of material inferences from one fact to another, there is often a certain degree of messiness and ambiguity that makes it likely that opponents will disagree about what inferences are valid—much more so than in a case of abstract reasoning like *if* $p$ *then* $q$, *p*, therefore $q$.

Nonetheless, the possibility of meaningful deliberation hangs on valid inferences having some sort of binding force on all participants. In ambiguous and complicated disputes, this requirement takes the form of a sort of good faith. That, as I argued in Chapter 6, is what is implied in public reasons: reasons are public where the people citing them consider themselves to be accountable to each other for their goodness. That means that while you might disagree with an opponent about the strength of many arguments, if a valid argument were presented, you would recognize it as such and abide by its conclusions.

But is that assumption of good faith well grounded? Unfortunately, in many cases it is not. The problem is that political opponents who would deliberate with each other care about where deliberations end up. Like Judges 2 and 3, rather than accepting and making arguments in good faith, people can be selective in their recognition and presentation of valid inferences. Where the premises do not lead to the outcome they want, they may either flatly refuse to engage in further argument or, if they want to keep up the appearance of deliberating, reverse the priority of premises and conclusions, “reasoning” backward from conclusion to premises—which is to say, they will rationalize.

---

37 Following the massacre at Sandy Hook Elementary School, the National Rifle Association made such an argument. Had the school principle had an assault rifle herself, she might have been able to stop the killer. As it was, she bravely attempted to tackle him unarmed. The gunman killed her and continued to kill others.
Rationalization can take many forms: you can ignore or cover up inconvenient facts; be selective about the conclusions you draw from a given piece of information or the way you interpret ambiguous information; frame an issue in a way that leads to your chosen conclusion; or engage in any number of fallacies (ad hominem, straw man, etc.). In addition, all the perverse incentives that pertain to revealing information apply here as well. Political opponents will have incentives to hide or misrepresent information that would serve as a premise for conclusions they wish to avoid.38

Whatever form it takes, this sort of backward reasoning, starting with the conclusions you want to emerge from deliberation and adjusting premises from there, subverts the process of deliberation. These incentive problems will arise wherever an outcome that is collectively rational—an outcome that makes sense according to inferences communicated in deliberation—remains disconnected from the reasoning that individuals find compelling. Incentive problems are an inevitable result of failures of collective will-formation. When individuals cannot relate to collective reasons as reasons—when they do not see collective reasons as their own—they will not be motivated by them. The incentive incompatibility of deliberation is simply the form that the disconnection between collective and individual reasons takes when those reasons are considered in their motivational, as opposed to their endorsement, aspect.

38 Landa and Meirowitz 2009.
7.3.3 A problematic answer: deliberative preferences

What options are there for escaping this motivational disconnection between collective and individual reasons? So long as we think of individual reasons as being private reasons, I don’t think there are any good ones.

One option is to hope that incentives to rationalize or mislead simply do not arise. This would work if people agreed from the outset on what conclusions should be reached—but then there would be no need to deliberate in the first place, and at any rate such prior agreement is absent in contexts of political conflict.

A second possibility is that adequate policing measures could be implemented so that rationalization is caught and punished firmly enough to discourage it. This option may be slightly more promising than the last: it is possible to imagine scenarios in which gaining a reputation for making bad arguments could harm a person’s ability to wield influence in the future, and thus serve as a disincentive to insincere deliberation. But such mechanisms, while they might occasionally have an effect, hardly seem up to the task of guaranteeing the integrity of deliberation in a great many cases where we appear to rely on it. Rationalization is often hard to detect precisely because it exists where standards of inference and lines of reasoning are ambiguous and contested. It is often difficult to say conclusively that someone is arguing in bad faith, as opposed to simply seeing the issues differently.

A third option is perhaps the most promising. It lies in positing that people have specifically deliberative preferences. The incentive problem arises because political opponents have, in effect, attached their preferences to argumentative outcomes or conclusions rather than to premises. But what if the reverse were true? What if they valued getting the premises right, and allowed the conclusions to follow from there? Participants whose preferences were
structured in that way would not face the temptation to falsify premises or misrepresent their implications because getting the argument right would be what they most prefer.

Such preferences are quite conceivable and, indeed, they are plausible to ascribe to people in many social settings. Such preferences would represent a “deliberative ethos”: a commitment to getting the facts right and following arguments where they lead.\(^{39}\) To have such preferences is, in effect, to trust in the process of reasoning with others, even where the results of collective reasoning lead to conclusions that, considered apart from the process of decision that led to them, you might not support. We would expect decision-makers in many situations to have such preferences, particularly in technocratic domains, such as law, administrative review boards, and scientific enquiry. We expect judges, scientists, and bureaucrats to be committed to correctly applying rules of decision-making, preferring a correctly made decision to one that conforms to their prejudice or personal preference for a particular outcome. Such professionals are socialized and selected to have such preferences, and it may even be that these preferences come to form part of their identities. A political scientist who identifies with her profession, and sees a commitment to rigorous method as part of being a political scientist, will not be tempted to “massage” her data to get the regression curve she wants.

Nonetheless, there are at least two problems with this preference-based solution to the motivational disconnection between collective and individual reasons: one that lies on the surface and another that is more fundamental. The more obvious problem is that while judges, scientists, and bureaucrats might be socialized to have deliberative preferences, it is much less likely that political opponents have such preferences, particularly in conflicts where there are valuable goods at stake. The sorts of shared norms and identities that underpin a preference for

\(^{39}\) List and Pettit 2011, 114.
impartial reasoning in professional circles, or even among lay citizens in well-structured deliberative forums, are conspicuously absent in many political conflicts.

There is, however, a more fundamental problem, one that raises doubts about whether preference structuring is the best way to understand the connection between collective and individual rationality, even in professional contexts. The problem is that positing such preferences does not really address the disconnection between collective and individual reasoning because an individual who merely has a preference for abiding by collective reasons does not really treat those reasons as reasons. The individual’s own judgments are not integrated with collective judgments in such a way that the individual recognizes the collective judgments as right. Rather, he simply defers to them. The collective reasons do not become his reasons because his adoption of them is not itself rational, which is to say, it is not an act of judging.

Look again at the example of the court deciding the breach-of-contract case. In order for the verdict to be structured by the appropriate legal reasoning, the judges have to allow their collective judgments on the premises to determine the conclusion. The court will then find (a) that the defendant had a contractual obligation not to take the action in question, (b) that he took that action, and therefore (c) that he is liable for the damages. For Judge 1, there is no problem. But for Judges 2 and 3 who, for different reasons, believe that the defendant is not liable, handing down the verdict of “liable” means contradicting their own judgments. They may agree to do so because they have internalized a deliberative ethos of deferring to majority opinion on the premises of legal arguments. But if they do so, they will still not, in their own individual judgment, accept the verdict as right. They might accept the group’s decision, but not because their own judgments have been integrated as judgments into collective judgments, such that
collective reasons are their own reasons; rather, they accept the group’s decision as an outcome, as a compromise or aggregate result of a process, but not itself as a correct choice.

I find this to be an unsatisfying conception of deliberation. If we want to understand deliberation as a process by which individuals reason with each other, and in which the force of good reasons shapes the results, then it seems incongruous to suggest that deliberation relies on the sacrifice of individual judgments to collective ones. To be clear, voting on premises and drawing conclusions from there is not an unsatisfying way of making collectively rational decisions in the face of persistent disagreement. Deliberative democrats recognize the value of voting as a democratic decision-making procedure. The point, rather, is that premise-based aggregation is not a very good description of deliberation itself. It appears to achieve collective rationality at the expense of individual rationality, which makes it difficult to see deliberation as a process of genuine collective will-formation.

7.3.4 A better answer: public reasons

I would suggest that there is a better way to understand the relation between individual and collective reasons in deliberation, and that is to understand individual reasons as public reasons. The discursive dilemma itself is, I would suggest, a product of the implicit assumption that individual reasons are held as private reasons, and that is the source of the apparent necessity of choosing between individually rational and collectively rational outcomes. The possibility of deliberative influence, of agreement wrought through the rational force of valid inferences,

\[\text{\begin{footnotesize}40\end{footnotesize}}\text{List and Pettit 2011.}\]

\[\text{\begin{footnotesize}41\end{footnotesize}}\text{E.g. Goodin 2008, chap. 6; Mansbridge et al. 2010; Warren 2012b.}\]
depends upon understanding practical reasons as public. It is public reasons that can be at the same time the individuals’ own reasons and the reasons of the group.

As I argued in Chapter 6, what makes a reason private or public is its normative reach, that is, the extent of the justificatory community to which it is addressed as a reason. A private reason is one that an actor considers binding for herself. It is, in other words, a reason she would cite in justifying her action to herself or to her own constituency. A public reason, by contrast, is a reason that an actor accepts as binding for a public, that is, for a group of actors that includes all the others who are subject to a given collective action problem arising from a problematic interdependency. As a result, endorsing a public reason involves placing yourself in a relationship of mutual accountability not only to yourself or your constituency, but also to the others with whom you are interacting.

Given this, we can see that within the terms of the discursive dilemma, individuals’ reasons are represented as private. That is clear from the way reasons behave in such models. Within the models, reasons (or judgments or preferences) are fixed. The moving parts of the models, what changes as people interact with each other, are elsewhere. In the problem of judgment aggregation, such as the “doctrinal paradox” faced by the judges in the breach-of-contract case, the “action” in the model lies in the various ways individual judgments can be aggregated, whether by a majority vote on individual propositions or, as in the “deliberative” option discussed above, by a premise-based procedure of voting on premises and reasoning to conclusions from there. The key point, however, is that those individual judgments themselves are not in flux. They are the fixed points around which other adjustments are made. Likewise, in game theoretical models of communication, the “action” in the model concerns the choice of strategies, including the truthful or less-than-truthful representation of information and
inferences to others. But each actor’s own reasons, now in the guise of preferences, are the fixed points around which other adjustments are made.

And that is where the difference between public and private reasons lies. So long as reasons (or judgments or preferences—whatever intentional states ultimately ground the individual rationality of the choice) are assumed to be fixed for the purposes of interacting with others they are, in effect, private reasons, because it is essential to public reasons that they are not fixed, but rather are open to discursive challenge and criticism from others. To address a reason to another person as a reason, as a validity claim, is necessarily to present it as being open to revision.

Once the reasons given in deliberation are recognized as public, the conceptual difficulties captured in the discursive dilemma can be resolved. The apparent necessity of choosing between individual and collective rationality is an artifact of conceiving of reasons in such a way as to exclude the possibility of their revision through the rational criticism of others. The root of the problem is precisely that the mutual accountability implied in publicity—the way that individuals are accountable to each other for the rightness of their reasons—is implicitly excluded when reasons are represented as fixed preferences or judgments.

Where reasons are public, though, it becomes possible for actors to influence not only each other’s strategies, but their judgments, motivations, and preferences, too. The fact that individuals see themselves as accountable to others for the rightness of their reasons is precisely what makes it possible for others to influence them. And where deliberative influence takes place the result is a reason that is genuinely shared—a reason that belongs to the group, the public, to whom it is addressed, and is at the same time each individual’s own reason for acting.
Thus, rational persuasion—made possible by the responsive opening of the individual’s reasons to the criticism of others—dissolves the need to choose between individually and collectively rational decisions. Shared reasons are the actor’s own reasons in both their judgment and motivational aspects. As a matter of judgment, where persuasion takes place, individuals endorse public reasons as right. At the same time, the individual is not motivationally disconnected from the public reason because he shares it as a reason—that is, as a compelling justification for action.

The limitation of most rational choice models of deliberation can be summarized thus: in implicitly representing reasons as private, and thus excluding the possibility that individuals’ reasons might be responsive to the rational criticism of others, they overlook the form of interaction that is most distinctive of deliberation. That form of interaction is, as the pragmatist semantics outlined in section 7.1 suggests, a practice of ascribing deontic statuses to each other (Brandom’s “deontic scorekeeping”), a practice of mutual accountability among speakers for the rightness of the claims and intentions they espouse. If the pragmatists are right that reasoning together is just such a practice, then we should not be surprised that models that implicitly rule mutual accountability out should struggle to account for how deliberation works.

7.4 The contingency of public reasons in communication

If the argument of this chapter is so far correct, deliberation has to be understood as the give and take of public reasons. This places relationships of mutual accountability at the core of deliberation. Deliberation is an essentially norm-regulated practice. The normative statuses that interlocutors extend to each other—the attributions of commitment and entitlement to claims and counter-claims—are not an incidental addition to deliberation. Rather, they are the vehicle of
shared inference itself. To the extent that inferences or reasons, and not only information, are communicated, that communication must be understood as being structured by practices of shared normative accountability.

One question this raises, however, is how we should understand the contingency of deliberative, as opposed to non-deliberative, communication. If deliberation does indeed depend upon normative accountability, it is just as clear that in at least some circumstances, words are used in ways that defy such accountability. Words can, of course, be deployed in purely instrumental ways. If they are instrumentalized surreptitiously, so that they maintain the outward form of sincere arguments, they can manipulate or deceive. And in environments where people expect words to be used anomically or instrumentally (that is, the sort of environments modelled in non-cooperative game theory), they may work as costly signals to convey information or, in the context of coordination problems, as cheap talk. In short, if deliberation depends on norm-governed interactions and mutual accountability, that is not true of all language use.

That raises a puzzle for the pragmatic philosophy of language I described at the beginning of this section. If language is originally the sort of normatively structured social practice described by pragmatists, how then should we understand seemingly non-norm-governed or anomic uses of language? What, in other words, is the nature of the contingency of normatively structured and instrumentalized language use?

Answering this question is, I think, important for understanding the political implications of the philosophical pragmatics upon which deliberative democratic theory is founded. I would suggest that there are, broadly speaking, two answers.

The first sees the contingency of instrumental and norm-governed communication as a matter of action type or the action orientation of individuals. On this view, people abide by the
normative commitments necessary to deliberation because they have acquired, through socialization, the capacity and disposition to follow social norms. Instrumentalized language use is a result of suspending that normative orientation: of shifting from a norm-conformative to an instrumental mode of rationality.

The second characterizes the contingency of instrumental and norm-governed communication slightly differently. It is a matter not of different forms of action but of different forms of interaction. When people engage in instrumentalized communication, they are not acting differently, rather, they are adapting to the presence or absence of relationships of mutual accountability among the people communicating.

I will argue that the second option is a better way to account for the contingency of norm-governed and instrumentalized communications that we see across different political contexts. But first it is important to see how a reading of the pragmatic philosophy of language, and the social theory that goes with it, might suggest the first option.

7.4.1 The essential normativity of speech and thought

One of the central accomplishments of Habermas’s work is to connect his pragmatic philosophy of language to a social theory. A key aspect of this social theory is a conception of socialization, the processes by which individuals are formed as social actors. Habermas, drawing on a long tradition of social theory including the work of Émile Durkheim, George Herbert Mead, and Talcott Parsons, argues that central to the process of socialization is the acquisition of

42 Habermas 1984; Habermas 1987; Habermas 1990.
specifically normative competencies, including the competency to act in accordance with norms.\textsuperscript{43}

Habermas’s primary contribution to this tradition of social theory is to give a detailed account of how communication is central to this socialization. Habermas argues convincingly that linguistic competency—the ability to use words meaningfully—requires the adoption of what he calls the “second person” or “participant” perspective of judging the validity of claims. In other words, linguistic competency demands the normative competency of holding oneself and others responsible. We acquire the norm-conformative disposition of socialized actors as a necessary part of becoming language users.

Brandom’s pragmatics underscores this point by emphasizing the even more foundational connection between language and thought itself. Pragmatism is an account not merely of language but of \textit{concepts}. It is not just the meaning of words that is at stake, but meaning itself: that is, conceptual or propositional contentfulness, whether those concepts are (at the moment) outwardly expressed or remain only as inward thoughts. Pragmatism purports to provide an account of cognitively accessible consciousness, or \textit{sapience}, itself.\textsuperscript{44}

This broad-reaching claim stems from an appreciation of the internal connection between thought and language. Thought, on this view, is simply language that is left unspoken or, perhaps, language spoken to oneself. In this way, the normative structure of linguistic \textit{social} practice is seen as the source of \textit{individual} sapience. We come to be thinking, concept-using beings by engaging in the practice of deontic scorekeeping with others.\textsuperscript{45} In this way,

\textsuperscript{43} See also Heath 2001a.

\textsuperscript{44} Brandom 2000.

\textsuperscript{45} Brandom 2000, 6;
pragmatism reverses the more common order of explanation between intentional states and language. As I explained in section 7.1 above, rather than explaining the meaningfulness of words with reference to the intentional states of the actors who utter them (a representational semantics), pragmatists explain the intentional states of actors—the conceptual content of their thoughts, desires, and so on—in terms of the proper use of such concepts in linguistic practice (an inferential semantics). Our thoughts have conceptual content because we know what other concepts they inferentially rely upon and imply, and these inferential proprieties are themselves an emergent property of the normatively structured practice of giving and asking for reasons.

From this perspective, how should we understand the phenomenon of instrumentalized uses of language? If we see norm-governed communication as the vehicle of socialization, that suggests that the instrumentalized use of language must be the result of unpicking the socialized disposition toward norm-conformative behaviour. While communication is essentially a norm-governed process, and while the acquisition of basic linguistic competencies—and indeed, rational competencies, as Brandom argues—can only be done by taking a normative attitude to oneself and others (Habermas’s “second person perspective”), once those competencies are acquired it is possible to instrumentalize them and use them in anomic ways. You can, so to speak, “kick the ladder out after you climb up.”

On this view, then, using words in non-norm-governed ways means changing one’s action orientation or way of acting. Rather than prioritizing normative over instrumental considerations, which is the basic competency acquired in socialization, you reverse the order, allowing instrumental considerations to prevail. If the pragmatists are right, it is not possible for

\[\text{46} \text{ Brandom 2000, 11–12.}\]

\[\text{47} \text{ Neblo forthcoming, 112.}\]
language use in general to be the result of purely instrumental attitudes; but once a normatively structured language is in place, words can be used “parasitically” to instrumental ends. Likewise, while the basic linguistic and rational competencies of individuals can only be acquired in a norm-conformative practice, once formed, those competences can be used to engage in instrumentalized, anomic social action.

7.4.2 The contingency of normativized interaction

I would like to suggest, however, a slightly different way of understanding the contingency of norm-governed and instrumentalized communication. On this view, the difference lies not so much in different ways of acting but in different contexts of interaction. When people shift from norm-governed to instrumentalized ways of using language, I suggest that they are not undoing, even selectively or temporarily, their basic norm-conformative dispositions and rational competencies. Rather, engaging in instrumentalized language use may often be an application of those dispositions and competencies, an adaptation to interacting with others in an uncertain social environment. An uncertain social environment is one in which the commitment of others to practices of mutual accountability—Brandom’s “deontic scorekeeping”—is absent or at least suspect. In situations like these, properly socialized actors, actors who adopt an attitude of normative accountability toward themselves and others, will have reasons to communicate (to the extent that communication is possible) in ways that do not rely upon mutual accountability.

The difference, in other words, comes down to the existence or non-existence of a public. Publicity, as I argued in Chapter 6, is a relational phenomenon. A public is a group of people

48 Habermas 1984, 1:288.
who stand in relations of accountability to each other, who take themselves, that is, to be accountable to each other in terms of reasons. Where people who are interacting—or, for the purposes of this chapter, communicating—form a public, that means that the relational prerequisites are in place for public reasons to be valid.

And that, I would suggest, is where the contingency of deliberative and non-deliberative communication lies. Deliberation is possible and desirable precisely where people understand themselves to form a public. Thus, in seeking to understand the conditions for deliberation we should look not so much at the level of individual socialization or action orientation—whether individuals are generally disposed to norm-conformative actions or not, whether they are “instrumentally rational” or “normatively rational,” whether they follow a “logic of appropriateness” or a “logic of expected consequences.” Rather, we should look to how they understand their social context and to what sort of relationships they see themselves standing in with each other.

One way of representing the shift from the first option I described to this second one is as a shift from social to political theory. In a social theory (at least in the tradition of social theory on which Habermas builds), a central problem is the normative integration of the individual into society. From this point of view, the alternative to integration is anomie: the lawlessness and egoism of the individual who does not feel the force of social rules. But in a political theory the emphasis has to shift. The central problem in politics is not how to integrate individuals into society, but rather, how to form a public, a group of people (or peoples) who have established the relationships of mutual accountability necessary to use speech, and not only coercion, to manage conflicts. In a sense, the political question picks up when people (who are assumed to be socially formed individuals with normative and linguistic competencies) begin to mobilize power against
each other. If the alternative to successful social integration is anomie, the alternative to the successful formation of a public is war—or at least what Kant called a state of war: a situation in which conflicting groups have not yet accorded each other the basic recognition of mutual responsibility that makes reasoning with each other a possible response to conflict. One of the benefits of thinking about public and private reasons as I have set them out here is that it turns our attention to this kind of contingency.

7.5 Conclusion

In this chapter, I have argued that the reasons at play in deliberation should be understood as public reasons. The reason-giving that is distinctive of deliberation as a form of communication is only possible if people relate to each other on the basis of mutual accountability. Where people treat reasons as private, communication is limited to the signalling of information. Thus, in any particular context, the possibility of deliberation depends upon the formation of a public. How a public is formed, and what role deliberation might play in that, will be taken up in the next chapter.
Chapter 8  A Theory of Deliberative Relationship-Building

This chapter begins the final part of the dissertation. The task of Chapters 5 to 7 was to develop some conceptual building blocks for a deliberative theory of rule-based cooperation. Chapter 5 examined *practical rationality*, or what it means to act for a reason. Chapter 6 examined *public reasons* and their connection to collective action. And Chapter 7 examined *deliberation* as a process of the give and take of public reasons. These three concepts—practical rationality, public reasons, and deliberation—are the core concepts we need to construct a deliberative theory of the emergence of cooperation in anarchy.

The task of this chapter and the next (Chapters 8 and 9) is to construct that theory: to turn the conceptual connections I have traced in the last three chapters into a theory of how rule-based cooperation emerges. To this point, I have developed a detailed conception of what *rule-based cooperation*, on the one hand, and *deliberation*, on the other, look like. Both are inherently connected to public reasons and mutual accountability. What we do not yet have is an account of where those relationships come from. That is to say, if people see themselves as accountable to each other, and vice versa, and thus treat their practical reasons as public, the result will be broadly cooperative solutions to collective action problems, and deliberative responses to conflict. But what we do not yet have is an account of where those relationships come from, or the processes by which they are generated and decay. The goal of this chapter and the next is to do just that: to cash the work of the previous chapters out into a theory of how productive, cooperative, and deliberative forms of politics emerge under relatively anarchical circumstances.

The broad framework of that theory can be represented as a tree diagram (Figure 8.1). At the top, *stable cooperation* is a combination of *consent* and *coercion*. By “consent” I mean
cooperation that is essentially voluntary: cooperation that can be understood as the result of collective will-formation. Conceptualizing this kind of cooperation is the central goal of this dissertation and, as I argued in Chapters 3 and 4, the main contribution that deliberative theory can make to understanding cooperation in anarchy. “Coercion” refers to the use of coercive power to backstop and support voluntary cooperation. Defining the relationship between consent and coercion, and the contributions each makes to stable cooperation, is the topic of Chapter 9.

Figure 8.1. Components of stable cooperation

At the next level down, consent branches into reasons and relationships. These are the two dimensions of collective will-formation I first distinguished in Chapter 3. “Reasons” refers to the content of collective will: the public reasons that opponents act upon together. By “relationships,” I mean the relationships of mutual accountability that underpin public reasons, collective action, and deliberation. As I have argued throughout this dissertation, public reasons are only compelling if there is, in fact, a public: a group of actors who see themselves as accountable to each other in terms of reasons. The relational dimension of collective will-
The plan of this chapter is as follows. The main focus will be on theorizing the relationship-building aspect of collective will-formation. I will start, in section 8.1, at the level of the actors’ attitudes: their commitments to public reasons, and their perceptions of others’ commitment (or lack thereof). Section 8.2 builds on this to conceptualize relational steady states: relatively stable combinations of perception and commitment that result in either cooperation or defection. Section 8.3 introduces the possibility of relational change, in which the basic interactions modelled in section 8.2 are interrupted by relationally generative moments of deliberation. Section 8.4 adds the possibility of feedback loops between deliberation and collective action. The result is a conception of how deliberative politics, understood as a combination of deliberation and collective action, can generate the conditions of its own success by thickening relationships of mutual accountability. But there are also corresponding vicious circles in which non-deliberative and coercive interactions progressively undermine mutual accountability. Finally, in section 8.5, I reconnect this account of deliberative relationship-building with the reasons dimension of collective will-formation.

8.1 Commitment and perception of the other

A good place to begin in developing a conception of how relationships of mutual accountability are generated and decay is by looking at the relationship in more detail, from the perspective of an actor in that relationship. From the actor’s perspective, a relationship of mutual accountability has two elements: the actor’s own attitude toward the other, and the actor’s perception of the
other’s attitude toward her. For convenience, I will refer to these two elements as *commitment* and *perception*.

By *commitment*, I mean commitment to being part of a public, as I have defined that in the preceding two chapters: commitment, that is, to the practice of mutual accountability. To be committed is to hold yourself to be responsible to the other in terms of reasons, and to hold the other to be responsible to you. Commitment is, in short, your willingness, so far as your part is concerned, to relate to the other on a basis of mutual accountability.

This commitment has ethical or moral content. To take a stance of mutual accountability toward the other implies a certain sort of respect. To think that you owe reasons to another is to imply that the other has a particular moral standing: the standing of an autonomous actor, of someone capable of judging the reasons you might offer, and someone with the right to offer reasons of her own. Where political opponents are not willing to be accountable to each other—something they demonstrate by not taking each other’s claims seriously, or not offering justifications for their own actions—they withhold that particular moral standing from each other. To fail to take oneself to be mutually accountable to another is to fail to treat the other as having either the capability or the right to make autonomous judgments (or both).

Political opponents may have many reasons, good and bad, for doing so. It is easy to think of bad reasons for denying one’s accountability to others, such as racial or ethnic superiority. But there can also be good reasons for withholding that standing from others, particularly when the other is a corporate actor. A government, for example, may be justified in refusing to take a rebel group seriously if that group is not, in fact, coherent enough as a group to
make judgments or keep commitments. Likewise, the right of a leader or regime to speak and judge for a state may be compromised if it carries out crimes against humanity.\(^1\)

As I argued in Chapter 4, the respect implied in a commitment to mutual accountability does not necessarily imply friendship or a shared identity.\(^2\) Such relationships can take place among people who “concede one another the right to remain strangers.”\(^3\) The respect implicit in mutual accountability is sometimes a grudging respect. Mutual accountability can work among people who don’t particularly like each other or identify with each other, but who nonetheless recognize each other as legitimate stakeholders and reliable, responsible, interlocutors.\(^4\) (This is sometimes suggested as a model for US–China relations.)\(^5\)

A commitment to mutual accountability also has, as part of its ethical content, an implication of equality. It is a specific sort of equality, though: an equality that rests on the way good reasons bind all. This sort of basic or formal equality might, however, exist alongside considerable inequality in other respects. That much is evident in the practice of international relations, in which there is a fairly broad consensus that “great powers” have a legitimate claim to a greater say—although they are still expected to be able to justify their actions to other states. Such an imbalance of power and resources can be compatible with relations of mutual accountability.\(^6\)

---

2 Mitzen 2013, 9–10; Erman 2013, 862.
4 See Kelman (2005, 646) on the idea of “working trust.”
6 Bull 1977, chap. 9.
Finally, the commitment to mutual accountability should be understood as conditional or provisional: it means “willing to be accountable to the other, as long as she takes the same attitude toward me.” Defining commitment this way allows us to describe an important phenomenon. In some cases—which I will discuss further below—political opponents are willing to uphold public reasons and mutually accountability but they do not act on public reasons because they do not believe the other is willing to do the same. In these cases, the basic attitude of respect, or at least readiness to respect, is present, but its realization is blocked because of a perceived failure of reciprocity. It is important that we be able to distinguish this situation from one in which the actor would not be willing to be accountable to the other even if she thought the other would reciprocate.

Having defined commitment and perception, and assuming for now that these are binary variables, we can recognize four possible combinations of attitudes (Table 8.1). I will list those combinations together with the types of actions they give rise to.
Table 8.1. Combinations of commitment and perception

<table>
<thead>
<tr>
<th></th>
<th>Perceive the other to be committed</th>
<th>Perceive the other not to be committed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Committed</td>
<td>Engage in deliberation and joint action</td>
<td>Defensively refrain from deliberation and joint action</td>
</tr>
<tr>
<td>Not committed</td>
<td>See other as naive; attempt to sucker by pretending to uphold public reasons</td>
<td>Cynically refrain from deliberation and joint action</td>
</tr>
</tbody>
</table>

First, if the actor is committed to mutual accountability and perceives the other to be as well, he will address public reasons to the other and act on those public reasons himself. In this situation, the actor sees himself as being mutually accountable to the other, which is to say that he sees himself as being part of a public. Thus, the relational prerequisite of the validity of public reasons is in place. As a result, it makes sense to engage in both deliberation and collective action on the basis of those reasons.

(Note: I am assuming here that the public reasons in question are materially good ones. In other words, the content of the reasons is not at issue: assume that the parties have a strong interest in achieving stable cooperation and can agree on a cooperative scheme. What is at issue is whether the relational prerequisites for those reasons are in place. I will re-introduce the question of the content of reasons—the reasons, as opposed to the relationships, dimension of collective will-formation—in section 8.5 below.)
Second, if the actor is committed but thinks the other is not, he will not engage in deliberation or collective action, although he will remain open to the possibility of revising this stance if his perception of the other changes. I refer to this as “defensive” defection. In this case, the actor holds a basic attitude of respect toward the other, but is forced to respond to the other’s lack of reciprocity by suspending his adherence to public reasons. The ambiguity of the actor’s stance here is captured, I think, in Kant’s position on international law. Kant recognizes that states remain in a “state of nature” vis-à-vis each other: in effect, they have not yet formed a public. As a result, international public reasons—including the standards of international law—are highly circumscribed and provisional. The result is that the rigorous moral principles that Kant applies in domestic life are greatly relaxed, particularly when it comes to the use of force. Nonetheless, even if the lack of a public limits the scope of obligations for now, Kant also suggests that states have a duty to try to change these circumstance by forming a public. The respect for humanity that is the bedrock of Kant’s ethics requires that states strive to change the very conditions that make a narrow set of obligations necessary.

Third, it is possible for an actor not to be committed to mutual accountability, and yet perceive that the other is. Such an actor would, in effect, view his opponent as naive or idealistic, a sentiment expressed, for example by Hitler to his inner circle about the British and Americans in the years before the Second World War. This situation offers the well-known possibility of “suckering” the other. If the other really is committed to mutual accountability, an actor who is

7 Kant 1991b, 171.
8 This is the rationale for the few rules that Kant lays down for international law, such as a prohibition of assassination. These things are prohibited because they “would make mutual confidence impossible during a future time of peace.” Kant 1991a.
not may be able to take advantage of that commitment, provided he is able to convince the other that he is committed, too.

Finally, it is possible for an actor to neither be committed to mutual accountability, nor perceive others to be. This possibility is captured in standard game-theoretical models. Both players act only on private reasons (i.e. preferences) and see themselves as accountable only to themselves and their own constituencies for the instrumental rationality of their choices. Moreover, they assume that others take the same attitude. As a result, such an actor will not take public reasons seriously in deliberation or abide by them in his actions, but neither would he attempt to persuade the other that he does abide by public reasons. I refer to this as “cynical” defection.

8.2 A baseline of stable, non-deliberative interaction

Once the possible combinations of individual perception and commitment are identified, we can begin to consider what happens when people who hold these attitudes interact with each other. My main goal in this chapter is to theorize deliberative interaction, but before I do that, it is useful to establish a baseline model of non-deliberative interaction. Such a model will provide a baseline in two senses. First, it will provide a conceptual baseline, a simplified image of what interaction looks like if deliberative possibilities are removed. In this way it will provide a starting point to which we can add complexity. Second, it will provide a description of a baseline of stable patterns of interaction—steady states—from which deliberation-driven change might take off.

To establish this baseline, consider a simple model with two players. Each player has one of the four possible combinations of attitudes outlined in the previous section and summarized in
Table 8.1: committed to public reasons and perceives the other to be too (C,P); committed to public reasons but thinks the other is not (C,~P); not committed to public reasons but thinks the other is (~C,P); or not committed to public reasons and thinks the other is not either (~C,~P).

The actions that each player takes are also described in the previous section and summarized in Table 8.1. For this simple model, though, I will lump all three species of defection together: defensive defection, cynical defection, and suckering. The reason to group these together is that in this model of interaction, each player can’t necessarily tell whether another’s defection is defensive, cynical, or an attempt to sucker them. Thus, the outward forms of action possible are reduced to two: cooperate (c) and defect (d). By “cooperate” I mean deliberating sincerely and following shared rules—that is, acting and speaking on the basis of public reasons. “Defect” means refraining from those activities, either because the player is not committed to public reasons herself, or because she thinks the other is not. If the player is both committed to public reasons and perceives the other to be as well (C,P) she will cooperate; in all other cases she defects.

Finally, the game consists of multiple rounds. Between the rounds, the players change their perceptions (P) of each other’s commitment based on each other’s action in the previous round. I will assume that they follow a simple pattern of inference: if the other cooperated in the previous round, they perceive the other to be committed to public reasons (P); if the other defected, they perceive the other not to be committed (~P).
Figure 8.2. Baseline interaction

<table>
<thead>
<tr>
<th>Player 1</th>
<th>Player 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C, P</td>
<td>C, P</td>
</tr>
<tr>
<td>C, ~P</td>
<td>C, ~P</td>
</tr>
<tr>
<td>~C, P</td>
<td>~C, P</td>
</tr>
<tr>
<td>~C, ~P</td>
<td>~C, ~P</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c,c) both cooperate (i)</th>
<th>(c,d) 1 cooperates; 2 defects defensively</th>
<th>(c,d) 1 cooperates; 2 sucks 1</th>
<th>(c,d) 1 cooperates; 2 defects cynically</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d,c) 1 defects defensively; 2 cooperates</td>
<td>(d,d) both defect defensively (ii)</td>
<td>(d,d) 1 defects defensively; 2 tries to sucker 1</td>
<td>(d,d) 1 defects defensively; 2 defects cynically (iii)</td>
</tr>
<tr>
<td>(d,c) 1 sucks 2; 2 cooperates</td>
<td>(d,d) 1 defect cynically; 2 defects defensively</td>
<td>(d,d) both try to sucker each other</td>
<td>(d,d) 1 tries to sucker 2; 2 defects cynically</td>
</tr>
<tr>
<td>(d,c) 1 defects cynically; 2 cooperates</td>
<td>(d,d) 1 defect cynically; 2 defects defensively</td>
<td>(d,d) 1 defects cynically; 2 tries to sucker 1</td>
<td>(d,d) both defect cynically (iv)</td>
</tr>
</tbody>
</table>

Notes: Outcomes in boldface are equilibria. Arrows represent movement from one round to the next. Outcomes in grey are mirror images of others with the positions of the players simply reversed.

The results of the interaction are summarized in Figure 8.2. The value of this model is that it suggests which combinations of commitment and perception might be stable and which will tend to be unstable. There are five equilibrium outcomes, set in boldface.\(^9\) To simplify

\[^9\] There is, in fact, a sixth “babbling” equilibrium, indicated by the double-headed arrow, in which the players alternate between 1(C,~P) 2(C,P) and 1(C,P) 2(C,~P). The players take turns cooperating and defecting defensively, as each goes back and forth between perceiving the other to be committed or not. I take this equilibrium to be an artifact of the simplified game structure rather than a plausible description of actual behaviour. A more plausible prediction is that this babbling equilibrium would resolve to either mutual cooperation, 1(C,P) 2(C,P), or mutual defensive defection, 1(C,~P) 2(C,~P).
interpretation, we can put aside some outcomes—including one equilibrium—as being mirror images of the others, with the roles of the players simply reversed; these are shaded in grey.

The result is four equilibria that correspond to relatively stable patterns of interaction, labelled (i) to (iv). One of these patterns involves mutual cooperation and three involve mutual defection. In the cooperative equilibrium (i), both players are committed to public reasons and perceive the others to be as well. Thus, both cooperate, and their cooperation confirms their perceptions of each other’s commitment.

In all three of the equilibria characterized by mutual defection, the players perceive each other not to be committed to public reasons. The three equilibria differ, however, in how accurate those perceptions are. According to the model, players interpret each other’s defection as evidence that each other is not committed to public norms, but that inference sometimes goes wrong because defection can be “defensive”: as I suggested in the previous section, even actors who are open to abiding by public reasons (and thus to deliberation and joint action) will still defect if they do not perceive the same attitude in their opponents. The three equilibria involving mutual defection thus differ according how many players are actually committed to public reasons.

In equilibrium (ii), both players are, in fact, committed to public reasons, but neither realizes this about the other, and as a result both defect. This pattern is similar to the equilibrium in a “Stag Hunt” game. Mutual defection here is, in effect, the result of a security dilemma. Each player misinterprets each other’s defensive defection as an attempt to “sucker” her, and thus defects in the next round. The result is self-confirming misplaced suspicion of each other’s intentions.

---

In equilibrium (iii), one player is conditionally committed to public reasons but the other is not. The player who is committed rightly perceives the other not to be, and thus defects defensively. The player who is not committed wrongly perceives the other not to be, and thus defects cynically. (By “cynical” defection, recall, I mean the defection of one who does not take public reasons or norms seriously and doesn’t expect others to either.)

Finally, in equilibrium (iv), neither player is conditionally committed to public reasons, and both have accurate perceptions of each other in that regard. Thus, both defect cynically: Neither takes public reasons seriously, nor do they expect each other to.

Each of these equilibria represents, I think, a plausible description of stable pattern of interaction, and a stable combination of commitment and perception within the actors involved. The combinations of commitment and perception that are not stable—and thus tend to move toward one of the equilibria above—generally involve one or both players attempting, successfully or unsuccessfully, to sucker the other. This is also an intuitively plausible result. We should not expect patterns of interaction and perception that allow people to be suckered to go on for long, especially among political leaders who are generally cautious and wary.

8.3 Deliberative possibilities for change

With this baseline model of stable combinations of commitment and perception in hand, we can begin to explore possibilities for how deliberation might bring about changes in these situations. The goal at this stage is to recognize plausible ways in which deliberation might bring about change in the perception or commitment of one or both players. If deliberation can indeed move participants from non-commitment toward commitment, or from perceiving the other as uncommitted toward perceiving her as committed, deliberation will have the relationship-
building capacities that are essential to the emergence of rule-based cooperation. In other words, my goal here is to argue that moments of deliberation can be morally or relationally generative.

8.3.1 Deliberative possibilities for changes in perception

The first way deliberation might contribute to a relationship of mutual accountability is by revealing a conditional commitment to public reasons that opponents had, but did not previously perceive in each other. In other words, deliberation might bring about a move from a pattern of mutual defensive defection—the pattern of equilibrium (ii) above\(^{11}\)—to mutual cooperation. In this case, deliberation is, in effect, correcting opponents’ misplaced distrust in each other.

Mitigating the relational damage done by disappointed expectations and self-confirming distrust is perhaps one of the most important things deliberation can do in conflicts. Political conflicts occur where standards of right and wrong are hotly contested. As a result, actions that one party takes to be justified are often viewed by the other as not only wrong, but obviously and culpably wrong—as thus as yet more evidence of an opponent’s treachery or depravity. As I argued in Chapter 6, when people are judging each other’s behaviour, the crucial distinction from a relational point of view is between dissent, good-faith disagreement between what, in fact, we owe to each other, and deviance, the deliberate disregard for obligations to each other per se.\(^{12}\) In a climate of suspicion, dissent can be mistaken for deviance, thus triggering a cycle of self-confirming distrust.

---

\(^{11}\) It is important, though, not to confuse this with a movement from one equilibrium to another within the model. Rather, adding deliberation changes the model itself. Combinations of commitment and perception that are equilibria in the simple model may or may not be equilibria when deliberation is added.

\(^{12}\) Heath 2001a, 152.
This, in effect, is what happens in the “security dilemma” of equilibrium (ii). Each person is conditionally committed to public reasons but defects defensively because she thinks the other is not. Each thus sees her own action as justified—and, given how each perceives the other, she is right. But according to the simple inference that drives the model, “if the other defected in the last round, she is not committed,” each interprets the other’s defection as evidence of her lack of commitment.

What might deliberation do? If opponents offer reasons for taking the actions they do, they may succeed in justifying their defection to each other and thus demonstrating that it was the result of disagreement or misunderstanding and not deviance. And this, in fact, is often what deliberation is about. People frequently argue about the rights and wrongs of each other’s actions. There is scope to do so because what constitutes “defection” is often ambiguous. Defection presumably means breaking the rules, breaking with a cooperative scheme or an agreed-upon way of acting. But rules and mutual obligations are rarely clear cut. Rules tend to evolve as new circumstances arise that were not anticipated before. They tend to contain implicit exceptions, “unless clauses” that justify a break with the letter of the law.\(^{13}\) If ostensible “defections” are in fact justified, then others would be mistaken to take them as evidence of deviance or a lack of commitment to public reasons. Offering reasons for actions opens the possibility of correcting this mistake. And likewise, by demanding explanations for offending actions, opponents concede that a legitimate explanation is at least possible—that their opponents might not, after all, be outright scoundrels and not beyond reasoning with. Offering or demanding reasons for an offending action thus holds open the possibility of building the social fabric of mutual accountability. Even where the party taking offense does not agree that the

\(^{13}\) Chayes and Chayes 1993; Chapman 2005.
other’s reason is a good one, it may be enough if the reason is recognized as one that could be offered in good faith.\textsuperscript{14}

This means, in effect, complicating the simple model above by adding some necessary nuance to the way players interpret each other’s actions and thus update their perceptions. The model relies on an overly simple way of inferring people’s attitudes, specifically their commitment or non-commitment to public reasons, from their actions. The inference “if the other defected then he is not committed” is simply incorrect when each player’s defection is defensive. By complicating this inference, deliberation offers the possibility of avoiding this trap and shifting toward an equilibrium characterized by mutually acknowledged commitments to public reasons, and thus cooperation.

But the ability of deliberation to complicate that inference also points to a darker possibility. Deliberation (or apparent deliberation) may also disrupt the inference “if the other defected then he is not committed” when that inference is correct, which would mean forestalling a defensive response to being suckered. In this case, seemingly legitimate deliberation might be used by a devious opponent to mislead a trusting person and thus take advantage of him. This is simply to make the well-known point that deliberation itself carries risks. That is precisely what makes anarchical situations tough cases for deliberation. I will return to this point later in the chapter.

\textsuperscript{14} Chayes and Chayes 1993.
8.3.2 Deliberative possibilities for changes in commitment

The second path for deliberative change lies in the possibility that deliberation might foster commitment to public reasons. Commitment, recall, refers to an attitude of basic respect toward an opponent, the respect implied in one’s willingness to engage with that person on a basis of mutual accountability. This commitment is conditional on perceiving that the other is similarly committed, and so it doesn’t necessarily result in what I have loosely termed “cooperation”: addressing public reasons to each other (deliberation) or acting on public reasons (joint action). Thus, someone who is committed to public reasons may still defect defensively, so to speak, but even as she does so, she maintains a willingness to take mutual commitments and shared norms seriously if she thought the other would as well.

The model above does not include any possibility of moving from non-commitment to commitment: it captures only changes in perception, and only according to the simple rule “perceive the other to be committed if and only if she cooperated in the last round.” On that assumption, two of the equilibria described—those in which one or both players are not committed—are hopeless cases for the development of relations of mutual accountability and, with that, rule-based cooperation and joint action.

But if it is possible for people who are not committed to public reasons to become so, then some of the worst cases of conflict might not be without hope. What would be involved in moving from non-commitment to commitment?

The essence of commitment is respect. To be committed to public reasons in one’s interactions with an opponent means seeing the opponent as someone who has the right and
ability to form good judgments, and as a source of value: in Rawls’s phrase, as a “self-authenticating source of valid claims.”

How might such a change happen? For one thing, in the context of political conflict, the change may be made easier by the fact that, although opponents do not currently relate to each other this way, each presumably has such relations with someone. This point returns to a theme of Chapter 7. The contingency of commitment to public reasons in political conflicts is not generally about acquiring the basic capacity to relate to others on a basis of mutual accountability. Acquiring that capacity, rather, is a part of the “primarily socialization” by which people become competent members of any society. Establishing mutual accountability in political conflicts is a matter of extending these practices to a group of people who were previously outside one’s moral community. Doing so may still be difficult, and it may require overcoming deep-seated prejudices, worldviews, or strategic cultures. But it does not require individual human beings to fundamentally change their way of being in the world or the nature of their rationality (see Chapter 5).

From this perspective, it is not implausible that political opponents could come to extend to each other the sorts of respect and moral standing that are required for mutual accountability. What is required, in effect, is the humanizing of the other: the recognition that the other is, in some way, “like us,” and by extension has the same capacities for judgment, action, and

15 Rawls 2005, 32.


17 In the case of corporate actors, though, it may require some reconstitution or remaking. This is a difference between individual human actors and corporate actors like states, armed groups, and bureaucracies. While the vast majority of individuals are constituted, through their basic socialization, to be able to deploy public reasons, I’m not sure the same is true of all corporate actors.
responsibility. Such humanizing moments can come in many forms. George Orwell, in *Homage to Catalonia*, gives a well-known example from the Spanish Civil War:

> At this moment a man, presumably carrying a message to an officer, jumped out of the trench and ran along the top of the parapet in full view. He was half-dressed and was holding up his trousers with both hands as he ran. I refrained from shooting at him…. I had come here to shoot at “Fascists;” but a man who is holding up his trousers isn’t a “Fascist,” he is visibly a fellow-creature, similar to yourself, and you don’t feel like shooting at him.\(^{18}\)

But such moments can also occur in deliberation. Jürg Steiner’s (2012, 72–74) account of the story told by a former member of the FARC in a discussion group that included former members of right-wing paramilitary groups in Columbia is an example (see Chapter 2). The speaker told the story of how he had come to join the FARC by being influenced as a teenager by a friendly recruiter, by being given opportunities to study, and by the contrast to the cruel treatment he received while in the army. Even though the story did not contain a formal argument, it had a justificatory purpose: it suggested even to those who were opposed to the FARC that the speaker was fundamentally like them, someone who was well-intentioned but pressed by difficult circumstances. As a result of this humanizing moment, the participants in the discussion began to take each other’s arguments more seriously—they became more committed, in other words, to public reasons and mutual accountability.

> When people justify themselves and offer reasons for what they do, they sometimes reveal a fundamental similarity, a shared humanity, that fosters a shift toward relationships of

\(^{18}\) Quoted in Walzer 1992, 140.
respect and reciprocity. Of course, that is not always the case. Deliberation can also reveal differences, including differences that opponents might find repugnant and dehumanizing. But in its capacity for self-disclosure, deliberation may sometimes effect a humanization of opponents.

But the relationship-building potentials of deliberation also go beyond self-disclosure. Because it involves addressing reasons to each other, the very act of deliberating requires participants to present themselves as being accountable to each other in terms of reasons. To deliberate, in other words, is necessarily to present yourself as committed. That much is inherent in the logic of reasons themselves. If a speaker were not bound by the rightness of his reasons—if he could disregard them on a whim—the presentation of those reasons would effectively be meaningless. Likewise, to take up reasons, to affirm or criticize them, is to present yourself as accountable to others in terms of reasons, otherwise the whole enterprise of criticism is meaningless.\(^{19}\)

Of course, it is possible for this self-presentation to be false. And if we are talking about “deliberation” among people who are not open or committed to mutual accountability, their self-presentations will be false. Such reason-giving is not really deliberation but only the outward form of it. However, it may be that even insincere deliberation could have the capacity to move opponents toward relations of mutual accountability.\(^{20}\) Here the mechanism of “social learning” and the internalization of social roles, discussed in Chapter 4, may come into play. Sometimes playing a role for long enough—addressing reasons to each other and thus pretending to be accountable in terms of reasons—may actually change the actor so that the role is no longer a

\(^{19}\) As Habermas (1987, 2:73) argues, “under the presuppositions of communicative action a hearer can reject the utterance of a speaker only by denying its validity.”

\(^{20}\) Warren 2006.
fiction. To engage in reason-giving without actually seeing oneself as bound by public reasons is hypocritical. And while we should be careful not to underestimate the human capacity for hypocrisy, hypocrisy can also be unstable. The mismatch between pretense and reality for people who engage in reason-giving without being serious about reasons creates a tension that wants to be resolved either in giving up the cynical charade altogether, as when talks break down and violence resumes, or in the actors themselves becoming less cynical.

Even if such transformations are difficult and unlikely, one of the strengths of deliberation as a means of conflict resolution may be that it offers at least the possibility of changing such attitudes, and thus of making new forms of collective action possible, precisely where nothing else can. In other words, if change on this level is difficult, that also means that when it happens, its impact on the interaction between antagonists is exceptionally important. And if any form of interaction can move participants toward seeing themselves as accountable to each other, deliberation is the most likely candidate, since it is, in effect, the practice of mutual accountability itself.

8.3.3 The relative difficulty of changing commitment and perception

Having surveyed some of the possibilities for relational change through deliberation, we may also be able to generalize about the relative difficulty of these changes. It seems reasonable to assume that of the two types, change in commitment is the harder one.

Non-commitment to public reasons in a given forum of interaction often reflects deep-seated attitudes toward the particular opponent, or toward out-group members in general. Non-

commitment is often rooted in a broadly pessimistic “realist” view of politics in general, and of anarchical circumstances in particular. Non-commitment may be a matter of a state’s or organization’s “strategic culture”: the basic set of assumptions it holds about the nature of political conflict. If state leaders, for example, see international relations as a dog-eat-dog world in which force is the only effective means, interests are fundamentally in conflict, and rules or norms do not apply—if those are the basic assumptions political elites and bureaucracies bring to conflicts—they will have a bias toward non-commitment to public reasons or mutual accountability with other states.

On top of the general attitudes encoded in strategic culture, non-commitment can also be rooted in fairly durable attitudes toward particular opponents. The respect inherent in thinking of yourself as being accountable to an opponent and vice versa is conspicuously absent in many conflicts. Where people see each other not so much as opponents or rivals but as enemies, the notion of mutual accountability can even be seen as repugnant. “We should justify ourselves to them?” Non-commitment can be rooted, as I suggested above, in perceived status hierarchies, whether based on race, ethnicity, or caste. Mutual accountability to a group that one perceives as inferior can be perceived as a loss of status.

By contrast, changes in how opponents perceive each other would appear to be easier. Such changes do not necessarily involve changing deep-seated attitudes toward political conflict in general or toward a particular opponent. If an actor is herself open to establishing relations of mutual accountability with the other, then a change in how she perceives the other requires only a belief that the other is “coming around,” that is, coming to see the world the way that actor


23 Petersen 2002.
herself does. The actor’s own commitment to establishing a public may even predispose her to relatively optimistic assessments of others’ intentions. People who hold commitments to reciprocity, rule-based cooperation, and talk as a means of dealing with conflict are probably more inclined to see others as being similarly committed—perhaps even, as realists in International Relations suggest, to the point of naiveté.

That said, a note of caution is also in order. Non-commitment to public reasons can be deeply rooted, but that does not mean that it cannot change. Indeed, as I will argue below, one of the most important strengths of deliberation as a form of conflict resolution may be precisely that it can effect changes at this level. At the same time, changes in perception don’t necessarily come easily. Even to opponents who are open to establishing relations of mutual accountability and public reasons, vicious circles of misplaced distrust can be hard to disrupt.²⁴

Nonetheless, as a general rule, perceptions of others’ commitment are probably easier to change than commitment itself. On that basis, we can also generalize about the prospects for change in situations where mutual defection appears to be entrenched. Three of the four “steady states” identified in the previous section involve mutual defection, but they differ in how many of the actors involved are conditionally committed to public reasons. At the more hopeful end is equilibrium (ii). Here both players defect defensively because they mistakenly perceive the other not to be committed to abiding by public reasons. This is a relatively promising situation because establishing a relationship of mutual accountability, and with that the possibility of collective action, requires only that they succeed in communicating their real attitudes toward each other. At the less hopeful end is equilibrium (iv). Here, neither player is open to entering relations of mutual accountability with the other. Establishing such relationships thus requires a change in

²⁴ Jervis 1976.
this fundamental attitude for both, as well as establishing the perception, on both sides, that the other has changed as well. Equilibrium (iii), in which one player is committed to establishing relations of mutual accountability and the other is not, is in the middle of the spectrum.

8.4 Deliberation and collective action: virtuous and vicious circles

The next step in theory-building is to add collective action to the mix. I will argue that there are complex and circular causal connections between deliberation, collective action, and relationships of mutual accountability. As a result, both virtuous and vicious circles are possible. In the virtuous circles, moments of deliberation help to build the relational foundations that make collective action possible, which in turn fosters further deliberation and collective action. In the vicious circles, by contrast, manipulative and coercive speech (the opposite of deliberation) undermines the relationships of mutual accountability needed for collective action, feeding into a corresponding downward spiral of coercive speech and action.

8.4.1 The circular dynamics of relationship-building

To begin, what reasons are there to suspect that deliberation and collective action exhibit circular dynamics, whether virtuous or vicious? I suggest there are a number of reasons.

The first has to do with the chicken-and-egg problem between deliberation and relationships of mutual accountability. I argued above that deliberation can contribute to generating relationships of mutual accountability by fostering recognition and commitment to public reasons, as well as the perception that others are committed, too. But it is just as clear that deliberation depends on relationships of mutual accountability. That is built into the very concept of public reasons, which are the vehicle of deliberation: the validity of public reasons depends on
the existence of a public. Addressing reasons to a political opponent only makes sense on the presumption that she considers herself accountable to you in terms of reasons. In short, deliberation requires trust; that is, it requires confidence that a relationship of mutual accountability is, at least to some extent, already in place. This suggests that the causal connections between deliberation and mutual accountability must be understood as circular: we cannot have mutual accountability without deliberation, but neither, it appears, will people engage in deliberation if they do not have some confidence in their mutual accountability.

The second reason to expect causal loops in deliberative trust-building has to do with the tight connections between deliberation and joint action. One implication of the arguments I have made in Chapters 6 and 7 is that both collective action and deliberation are intimately tied up with public reasons. To act together is to act on public reasons (Chapter 6). And to deliberate is to address reasons to each other as public reasons (Chapter 7). In this light, it is clear that deliberation and joint action are closely connected. Indeed, it may even make sense to think of them as an amalgam, as a single activity. Acting together on public reasons is only possible if there are such reasons available, and deliberation is the only way for public reasons to be established. Likewise, addressing public reasons to each other, responding to criticisms, and demanding justifications are all themselves actions—speech acts—and such linguistic actions are often part of collective action. When states conclude an agreement or issue a joint declaration, they are both acting together and deliberating at the same time. True, speaking together and acting together are not quite the same. Action sometimes has a finality or gravity that speech lacks; acting means “putting your money where your mouth is.” But the close connections between deliberation and joint action strongly suggest a causal loop: acting together tends to foster better deliberation, and better deliberation fosters acting together.
The third reason to suspect a causal loop is that joint action itself seems to stand in a similar chicken-and-egg relationship to mutual accountability as deliberation does. That suggests that even though the broad purpose of this dissertation is to develop an account of the role of deliberation in the emergence of collective action and rule-based cooperation, collective action is not solely a dependent variable, just as deliberation is not solely an independent variable. Collective action is not only the result of the deliberative development of public reasons and the relationships that underpin them. Collective action also contributes to those relationships. When political opponents act on public reasons—particularly when doing so means privileging a public reason over a private one, as happens when facing collective action problems—they affirm and enact their commitment to public reasons and to relationships of mutual accountability.

8.4.2 The virtuous circle

The way to accommodate all these observations is to recognize deliberation and collective action as activities that both depend on and contribute to relationships of mutual accountability (Figure 8.3). Deliberation and collective action, in effect, can be part of a virtuous circle. Moments of deliberation in negotiations between opponents help to build the relational foundations for further deliberation and for collective action. And when opponents act together—even if initial steps of cooperation are small—their actions do the same thing: they thicken the relationships needed for further cooperation and further deliberation.
We can refer to this combination of deliberation and collective action, of talking together and acting together, simply as \textit{deliberative politics}. Deliberative politics is what results when the relationally generative capacities of deliberation, which I surveyed in section 3 above, are strung together into a continuous activity. Deliberative politics can, in effect, feed on itself by creating the conditions of its own success. It not only draws on prior normative commitments and relationships; it also generates and thickens new ones. The virtuous circle between talking together, acting together, and the relationships of mutual accountability—the publics—that underpin both these activities is what allows political action based on talk, consent, and compromise to expand.

Developing a better understanding of this generative capacity of deliberative politics is important in at least two ways. First, it is necessary for understanding how deliberative politics takes hold under difficult circumstances—including relatively anarchical circumstances. In such circumstances, deliberation and joint action face an uphill battle: insecurity is rife, mutual distrust prevails, shared normative commitments and identities tend to be limited, and supportive institutions are weak. The same goes for war-torn or post-conflict situations, and for weak or failing states. In situations like these, the conditions for deliberative politics would appear to be absent. But if deliberation can begin to generate those conditions, and if it has self-stabilizing mechanisms, it is possible for talk-based politics to get a start even here.
But the normative generativity of deliberative politics is also important even in the more favourable conditions of stable democracies. Even here, citizens’ commitments to a politics based on public reasons requires constant renewal. Without the ongoing regeneration of the normative commitments and practices on which it rests, deliberative politics suffers erosion. Antagonistic partisan identities can harden, trust and mutual respect can unravel, and people can become cynical about political talk and action. Normatively generative mechanisms are needed for the ongoing maintenance of deliberative institutions as much as for their initial development.

8.4.3 The vicious circle

The existence of these feedback loops between deliberation, collective action, and mutual accountability should not, however, be taken as a teleological claim that the development of deliberative politics is inevitable. To the contrary, these feedback loops can also run in reverse: they also explain normatively degenerative politics, the erosion of publics, and the unravelling of shared rules and cooperation. Downward spirals of conflict take place as relationships of mutual accountability fall apart. Thus, understanding the connections between deliberation, collective action, and mutual accountability offers insight into both good and bad scenarios.

Figure 8.4. Vicious circle
The vicious circle by which publics fall apart is the mirror image of the virtuous circle by which publics are built (Figure 8.4). Both deliberation and joint action, which are normatively generative, have normatively degenerative counterparts. These degenerative activities undermine mutual accountability by either undermining opponents’ commitment to public reasons or their perception of each others’ commitment. Like the generative activities, they both depend on and contribute to mutual accountability, only now it is the opposite: they depend on and contribute to the erosion of mutual accountability. Thus they contain the same circular dynamics, only working in reverse.

The degenerative counterpart of joint action is defection, whether it takes the form of a cynical disregard for public reasons and obligations to the other or, worse, an attempt to use another’s trust against her. The potential for a vicious circle is clear. The victim of such defection will quickly downgrade her own commitment to shared rules, which will be reflected in her response. The result is the familiar pattern of “unravelling” as shared norms crumble and violence or bare power is used with fewer and fewer restraints.

The idea that defection prompts further defection is, of course, well-known, but its relational significance is often overlooked. From the perspective of public reasons and collective will-formation, the problem with defection is not just that it harms others, but that in disregarding public reasons, it defies mutual accountability. It is the undermining of mutual accountability that renders a coercive response from others, i.e. further defection, necessary: coercion is necessary and appropriate in situations of strategic interdependence in the absence of shared rules. But the fact that the victims of defection sometimes respond by demanding explanations indicates that it is not only the bare act of defection itself they are concerned with, but also what it indicates about the offending party’s commitment. The demand for explanation
is an attempt to sort out dissent from deviance—categories that indicate the relational significance of defection.

The degenerative counterpart of deliberation is manipulative or deceptive talk. What makes such talk a problem is precisely its relational significance: lies and manipulation undermine mutual accountability. This is perhaps most clear in the case of straightforward deception. Lying undermines the ability of others to hold me to account for the goodness of the (ostensibly) public reasons I claim and act on. But the same is true of more subtle forms of manipulation and framing through talk. To the extent that these undermine the judgment of the person they are addressed to, they undermine the relationship of mutual accountability on which meaningful communication relies. If my attitudes toward another person’s claim are in part the result of manipulation, I am not really able to evaluate those claims, and the accountability between us is compromised.

The result is a specifically talk-based form of normative unravelling. When opponents no longer trust each other to use words truthfully and openly, what unravels is the ability of words to convey meaning. When opponents can no longer rely on public reasons as shared standards of inference, communication is reduced to bare signalling and cheap talk.

8.5 Reasons and relationships in collective will-formation

The final theory-building step is to integrate the reasons and relationships dimensions of collective will-formation. In this chapter so far, I have focused on the relational side: on how relationships of mutual accountability are generated (or, as I pointed out in the last section,

decay). In order to form a complete picture of collective will-formation or consent-based cooperation, the relational dimension has to be put together with the reasons dimension (Figure 8.1).

To review briefly, I introduced the distinction between the *reasons* and *relationships* dimensions of collective will-formation in Chapter 3. Deliberative democrats most commonly think of collective will-formation in terms of reasons. Reasons are ostensibly what deliberation is about: deliberation can be defined as influence through the give and take of reasons. And when political actors succeed in persuading each other this way about practical questions, the result is collective will and collective action. Here again, reasons lie very much on the surface. To act together is to act on the basis of shared, public reasons.

But a central claim of this dissertation is that collective will-formation also has a relational dimension. Here the question is not “what is a good reason?” or “what reasons should guide our actions?” Rather, the relational question is “to what extent are we (the people deliberating) responsible to each other in terms of reasons at all?” In many ways, this relational dimension is the more fundamental one because the whole practice of giving and asking for reasons only makes sense among people who think of themselves as being accountable to each other. Thus, collective will-formation is not only about the *content* of collective will (i.e. what reasons the actors will agree upon as valid); it is also about *the constitution of the public*, the collective, as the kind of entity that can reason and act together in the first place.

What I have said to this point in the chapter is precisely about that: the constitution of the public. I have bracketed the question of the content of public reasons by assuming that there are public reasons that would prescribe broadly cooperative actions. But the existence and recognition of such reasons cannot be taken for granted. In a process of collective will-formation,
reasons and relationships must be built at the same time, and they are built through the same processes of deliberation and joint action.

The conceptual distinction between these dimensions can help to diagnose the limiting factors for the emergence of cooperation in cases of conflict or negotiation. In some cases, the extent of cooperation is determined by the availability of public reasons, that is, by the content of reasons that all parties can buy into. It is possible for negotiations to fail, or to achieve relatively limited success, because there is little the parties can agree to. Parties may simply be far apart in the reasons they find compelling. To some extent, these reasons can be understood as interests, and the distance between the parties as the extent to which their interests overlap. Reasons can, of course, be goal-oriented or teleological in content, and thus the extent to which the parties’ goals are compatible or incompatible will affect the extent to which they share reasons. Common interests will tend to create practical reasons that parties can share.

But even where interests, and thus “first order” reasons, conflict, parties may sometimes be able to find agreement on the basis of various forms of “second order” agreement, including “incompletely theorized agreements,”26 meta-consensus,27 principled compromises,28 and fair bargains.29 One of the important developments in deliberative theory has been to come to a more expansive conception of what counts as a reason, and of the types of rationally motivated agreement that are possible under conditions of value and interest pluralism.

26 Sunstein 1995.
27 Dryzek and List 2003; List 2007; Dryzek and Niemeyer 2010.
28 Gutmann and Thompson 1996; Bohman 1996.
29 Rawls 2005.
In other cases, though, the limiting factor is not so much reasons as relationships or trust. Here, the parties simply do not see themselves as being accountable to each other, or they do not trust each other to abide by public reasons in the face of conflicting private reasons. This may occur even though (from an observer’s perspective at least) there is considerable overlap in their goals and commitments and thus considerable scope for rationally motivated cooperation.\(^{30}\) In these cases, collective action is stymied not for the lack of a rational basis, but because the relational underpinnings of those reasons is not in place.

In fact, relationships may often be the real problem even when it appears that reasons are the limiting factor in cooperation. This happens when opponents don’t trust each other enough in order to undertake the sort of deliberations that would discover or establish shared reasons. Again, this applies to reasons that take the form of straightforward interests, as well as to forms of “second order” agreement.

Deliberation may bring new common interests into view in at least three ways. First, opponents can discover pre-existing shared interests by revealing their interests to each other more fully. Second, deliberation might also result in the discovery of productive solutions to problems, that is, ways of creating new common value that neither saw before. Third, deliberation may transform interests or preferences themselves, and in that way increase the scope for shared reasons.\(^ {31}\) But all these paths by which deliberation brings new interests into view may be blocked if the relational basis of deliberation is weak. When that happens, it may appear that there is “no point in talking”: the parties’ values are too apart and their interests are simply opposed.

\(^{30}\) Warren and Mansbridge 2013.

\(^{31}\) Landwehr 2009.
A full picture of collective will-formation, then, must involve both the reasons and relationships dimensions. In practice these go together. Political opponents build relationships of mutual accountability precisely by engaging in the practice of holding each other accountable through the give and take of reasons, and by acting together on the reasons for which they are responsible.

8.6 Conclusion

In this chapter I have brought together ideas from the previous chapters to build up a conception of how relationships of mutual accountability are generated through the combination of deliberation and collective action. Relationship-building takes place through complex and circular causal processes, in which talking together and acting together both draw upon and build the relational resources for further talk and action. These relationship-building dynamics provide an important counterpart to the shared reasons that are more often the focus on deliberative democratic theory. The end result is a complex picture of collective will-formation in which political opponents find and establish shared reasons to govern their actions, together with their growing commitment to, and confidence in, the public as a collective entity capable of acting together.

In the next chapter, I will take the final theory-building step of putting together this essentially voluntary conception of action with the building of coercive institutions to achieve a conception of institutionally supported cooperation.
Chapter 9 Will and Coercion in Institutions for Collective Action

Talk of institutions has been notably absent from the deliberative conception of collective action I have been developing. That absence, I think, is justified. The goal of this study has been to develop a deliberative conception of the emergence of cooperation in relatively anarchical, that is, institutionally thin, circumstances. The theory, thus, needed to be specified in such a way as not to depend on the prior existence of strong institutions. Instead, I have sought to frame the theory so that it could capture the sort of normatively generative interaction on which institutions and institution-building themselves depend. Part of the challenge this dissertation took up was precisely to explain how the sorts of shared normative commitments that underpin institutions are built. Institutions themselves are political creations; they are the result of political actors, who often see themselves as opponents, acting together. If we are to understand the generation of new institutions, and perhaps even the maintenance of existing ones, we need to understand those extra-institutional, interactive moments of politics in which the collective will to make and maintain institutions is forged. ¹ These are things that happen, so to speak, “‘upstream’ of law and policy.”²

But with that conception of will-formation in place, we now need to return to the subject of institutions. There are two reasons to do so. The first is that, as I argued in Chapter 3, stable cooperation requires the support of coercive sanctions to contain opportunistic defection. That is an enduring point made by Hobbes. In the absence of effective enforcement, cooperation is easily undermined by mutual fear and mistrust. Institutional guarantees are an important part of

¹ Arendt 1965; Honig 1991.
² Warren 2012a, 6.
giving people the confidence to trust and to engage in the voluntary cooperation and deliberation I have been describing to this point.

The second reason to discuss institutions comes directly from the central concern of the dissertation: cooperation in anarchy. In relatively anarchical circumstances, one of the main things people try to do when they act together is to make their interactions less anarchical. In other words, collective action is often about building institutions.

But institutions and institution-building pose some conceptual difficulties for deliberative theory. The problem is the place of coercion in institutions. Coercion is a central part of institution-building, particularly when it comes to institutions for collective action. Institutions facilitate collective action in large measure by organizing coercion, whether that takes the form of informal sanctions imposed by actors (or networks of actors) on each other or centralized authorities with the power to impose punishments. Sanctions are an essential part of the way institutions incentivize cooperation and discourage opportunistic defection.

To be sure, coercive sanctions are not the whole story. Institutions can also foster cooperation that is essentially voluntary. Institutions can give people good reasons to cooperate. One way they can do so is by facilitating productive bargains, mutually beneficial “win-win” solutions that all parties can willingly support. Institutions can facilitate such solutions by lowering transaction costs, providing structured forums for negotiation, or linking different areas of negotiation so that mutually beneficial bargains can be made. Furthermore, institutions may facilitate voluntary cooperation by transforming interests. This happens, for example, with international trade regimes when the lowering of trade barriers creates constituencies within states of export-oriented manufacturers. Such transformation of interests can create institutional

“lock in” effects and other forms of path-dependence that reinforce and expand voluntary cooperation.4 Finally, institutions can also foster voluntary compliance by embodying legitimate decision-making authority. To the extent that people buy in to the legitimacy of institutions, they will be inclined to obey them.5 In all these ways, institutions can foster willing compliance by giving people good reasons to cooperate.

Nevertheless, coercion also plays a significant role in institutions for collective action. And that fact poses the challenge I want to take up in this chapter: the challenge of putting the essentially voluntary or consensual conception of cooperation I have developed in this dissertation so far together with the role of coercion in institutions for collective action. As I suggested in Chapter 4, deliberation is essentially non-coercive: deliberative theory concerns the realm of normative commitments, persuasion, and willing cooperation. It is important, however, to relate these voluntary aspects of cooperation to coercion. I will argue that stable patterns of institutionalized cooperation combine voluntary and involuntary cooperation, collective will-formation with collective coercion.

The argument will unfold as follows. In section 9.1, I begin by distinguishing between voluntary and coerced cooperation. This is an essential distinction for any deliberative conception of rule-based cooperation. Section 9.2 presents an account of the role of coercion in stabilizing cooperation. I argue that even in cooperative schemes based primarily upon voluntary compliance and collective will-formation, coercion plays a necessary supporting role by incentivizing and backstopping cooperation. Sections 9.3 and 9.4, finally, consider more complex interactions between voluntary and coerced cooperation. On the one hand, coercive

5 Tyler 2010.
inducements can affect voluntary commitments to public reasons through “spillover” and “crowding out” effects. On the other hand, collective will-formation affects coercive inducement through institution-building.

9.1 The difference between will and coercion

A deliberative conception of cooperation and institutions must begin by distinguishing between voluntary and coerced cooperation. That distinction is essential to deliberative theory. Deliberative theory is about reasons and, in the realm of practical affairs, about will. Willing is an act of practical reasoning. To will an action is to act for some purpose, some reason, that you endorse as good. Thus, a specifically deliberative conception of collective action, such as I have developed in the last four chapters, has collective will-formation at its centre. This yields a conception of collective action as being essentially voluntary: it consists in the knitting together of people’s wills through the binding power of valid reasons.

Coercive power is fundamentally excluded from the concept of rational will-formation in deliberative theory. One litmus test for deliberation is the absence of coercion: that is the way to distinguish deliberation from other talk-based forms of influence, including coercive bargaining, deception, and manipulation. A deliberatively achieved agreement is by definition one that people buy into as good, one that does not depend upon force or threats to achieve compliance.

Despite its importance to deliberative theory, however, the distinction between will and coercion is not very well known or commonly used by political scientists generally. Other analytical distinctions are more common. We more often speak in terms of self-interest versus altruism, (material) interests versus norms, or the logic of expected consequences versus the logic of appropriateness. But from the point of view of deliberative theory, it is the distinction
between voluntary action and coercion that matters most, because that distinction is crucial to
marking the boundary between deliberative and non-deliberative influence.

The distinction between will and coercion, though, can be a tricky one to make,
particularly in the face of conflicting interests and value pluralism. The tricky part, I think, lies in
how we understand the inevitable constraints that each person’s interests impose on others. In
situations of interdependence, other people’s actions are often a constraint on my own. I cannot
simply do whatever I would like, because what I do will affect how others react, and their
reactions will in turn have an impact on me. It is nearly impossible for actors in situations of
interdependence to be wholly unconstrained by the actions of others. People who act as if others
do not exist will sooner or later be brought up short by the unfavourable reactions of others.

Thus, to be constrained, in some sense, by the actions or attitudes of others seems
endemic to social interaction. But does that mean that all interactions are essentially coercive? It
might seem that way. The very notion of constraint seems to imply an unfreedom, an imposition.
When an actor modifies her choices or claims in order to avoid the opposition of others, and
where those others have the capacity to harm her (as they do in a situation of interdependence), it
is hard not to see an element of coercion, an implicit “or else...” On this view, social interactions
are always underwritten by that implicit threat. The accommodations social actors make to one
another are all, to some degree, extorted, though perhaps the extortion is mitigated by the fact
that it is reciprocal. On this view, all social action is a version of “the gunman situation.”

----------------------

6 Chambers 2009b, 360.

7 In situations of extreme power imbalance, it may be possible for a powerful actor to act without regard to others,
but then this is not a situation of interdependence, but rather of the independence of one actor and the dependence of
others.

8 Cf. H.L.A. Hart’s (1961) comment that legal positivism amounts to seeing legal authority as “the gunman situation
writ large.”
To assimilate all constraint to coercive constraint, though, is a mistake. Not all constraints are coercive because it is possible for people to accept constraints placed on them as reasonable. This is the insight offered in the liberalism of John Rawls and others. Liberalism offers a conception of how it is possible for people to accept the inevitable constraints involved in cooperating with others across fundamental conflicts of interests and values—and, most importantly, to accept those constraints not as impositions, but as right or reasonable. The crucial moment of liberalism lies in taking others’ interests and values as presumptively legitimate, such that their values, like yours, should shape the common rules by which you live and cooperate. In that case, you will constraints on your own action, which is to say that you exercise self-restraint. And if you will those constraints, you are not coerced. In Rawls’s terms, these constraints are part of “ideal theory”: they are part of how a fair scheme of cooperation should work given the opposing interests and values involved.\(^9\)

What makes for coercion, then, is not the mere fact that others’ interests adversely affect your own. That, unfortunately, is inevitable in nearly any scheme of cooperation. Rather, what makes for coercion is the way your interests are adversely affected; in particular, that those effects are unilaterally imposed on you by someone else. What makes a sanction coercive is that it is against your will. The distinction between consent and coercion thus cannot be made in the language of interests alone; it also requires the language of will.

---

\(^9\) Rawls 1999b; Rawls 1999a. See also Kant 1991c.
9.2 Moral core, coercive periphery

If the deliberative theory of cooperation is concerned essentially with voluntary action, what place could there be for coercion? On the face of it, it is not clear that there is a place for coercion in such a theory. The arguments of pluralists like Rawls and others suggest that conflicts of interests do not in themselves necessitate relations of pure power. Contrary to the realist strain of political thought (from Hobbes and Machiavelli to Schmitt), it is possible to respond to conflict in fundamentally non-coercive ways. Even in the face of conflicting values and interests, cooperation can take the form of voluntary self-restraint on the basis of moral insight, and not simply a temporary truce in a battle for power or an unstable modus vivendi.

But this purely voluntary conception of cooperation is, I think, an incomplete one—at least, it is incomplete as an account of stable collective action among political opponents. Purely voluntary cooperation might be possible in some small, tightly knit social groups, where individuals are bound together by common interests and strong bonds of affection or identity—though even here it seems doubtful that interactions are ever wholly free of coercive power. Be that as it may, it is clear that among actors whose interests conflict and who have incentives to defect from shared rules, stable cooperation generally requires coercion, too.

The model I propose is what Jane Mansbridge calls a “moral core, coercive periphery” solution to collective action problems. In such a model, most people, for the most part cooperate voluntarily. That does not mean that those people do whatever they want. To the contrary, collective action often requires some degree of self-restraint, the choosing of public

\[10\] Habermas 1996, 121.

\[11\] Mansbridge 2001.
reasons in spite of temptations to the contrary. Furthermore, the fact that interests conflict means that shared rules and collective actions will rarely be anyone’s first choice. What makes collective action voluntary is that people accept those constraints and act on public reasons—that is, make them the object of their wills.

But in addition to this “moral core,” cooperation also requires some coercion “around the edges.”¹² What exactly that coercion does, and how it interacts with voluntary or rational motivations, is different for different actors, depending on how committed they are to upholding public reasons. But in the system of interaction as a whole, mutual coercion plays a necessary, though supporting, role.

The goal of this section is to explain the roles of coercion and consent in stable systems of cooperation. I will argue that in a system of action with a “moral core, coercive periphery” structure, coercion plays two roles: incentivizing and backstopping cooperation (Figure 9.1). The distinction between incentivizing and backstopping turns on whether cooperation is motivated primarily by private or public reasons.

---
9.2.1 Incentivizing cooperation

The first way coercive sanctions can support cooperation is by directly incentivizing it. This is the basic logic of using sanctions as deterrence: people can dissuade each other from cheating by threatening to impose punishments that outweigh the benefits of cheating.

Sanctions work by giving the potential cheater a *private* reason to follow the rules. Notice, first, that sanctions work by affecting the reasoning of the person they are aimed at. They do so by unilaterally changing the outcomes of the other’s actions, which, in effect, changes the environment in which she chooses. If cheating, which would otherwise be lucrative, now carries with it a significant chance of punishment, the actor now has a reason not to cheat, and that reason is the result of the credible threat made. (Figure 9.2b depicts coercive incentivizing, where A is the punisher and B is the potential cheater. Figure 9.2a depicts deliberative influence for comparison.)

Second, this reason, though, is a private one. Notice that the punisher does not address this new consideration to the other *as* a reason. The punishment is not presented to the potential
cheater as valid claim—or if it is, it is not its validity that makes it effective. (Otherwise the target would be persuaded not to cheat, not coerced.) Rather, the punishment is presented as a fact.\textsuperscript{13} That the potential cheater will have reasons to adjust to this fact is known and counted on by the punisher, but those reasons are treated by both as the cheater’s own, private reasons. There are, in effect, “normative silos” insulating the reasons of the punisher and the cheater so that there are no links of normative accountability between the two (represented by the crossed-out line in Figure 9.2b).

Figure 9.2. Deliberative and coercive modes of influence

\textit{a. Deliberative influence}

\begin{center}
\begin{tikzpicture}
\node at (0,0) (A) {A};
\node at (2,0) (B) {B};
\node at (1.5,0) (Public reasons) {Public reasons};
\draw[->] (A) -- (Public reasons);
\draw[->] (Public reasons) -- (B);
\end{tikzpicture}
\end{center}

\textit{b. Coercive influence}

\begin{center}
\begin{tabular}{c|c|c}
\hline
\multicolumn{3}{c}{Outcome of B's choices} \\
\hline
Private reasons & \texttimes & Private reasons \\
\hline
A & "Normative silo" & B & "Normative silo" \\
\hline
\end{tabular}
\end{center}

The fact that coercive sanctions work on private reasons means that they do not have the relational prerequisites that deliberation and joint action have. There is no need for actors to form

\textsuperscript{13} Habermas (1996, chap. 3) refers to this as the “facticity,” as opposed to “validity,” of law.
a public in order to influence each other through the force of threatened sanctions. Thus, not surprisingly, coercive influence predominates in “relationally thin” contexts of low-trust, anarchical interactions.

However, even though such sanctions do not depend on public reasons, they can work *alongside* public reasons. How they work depends on whether, and how firmly, the person being coerced is committed to public reasons. In a “moral core, coercive periphery” system, most people have significant commitments to public reasons. For these people, the private reasons provided by the threat of sanctions are an occasional and peripheral motivational contributor to cooperation. These people find public reasons rationally compelling and thus motivating, but they are also torn: they also feel the motivational pull of other considerations. In other words, they are tempted to cheat. According to the conception of practical reason I presented in Chapter 4, people in this position are struggling to integrate the various inclinations, incentives, and goals they experience into a will: a course of action they can endorse as good. Even where they recognize the force of public reasons, they may fail in this task. Here, the addition of potential punishments can help to counteract the motivational pull of other considerations, and thus tip the balance toward actually doing what they, in their better moments, recognize as the best choice.

This possibility complicates the purely voluntarist conception of action I have presented up to the start of this chapter. Where an actor’s cooperation requires coercion—even if that coercion contributes a relatively small part of the motivation—the cooperation is not wholly voluntary. Some element of the non-rational and the unwilled is present in the action, and thus the processes of collective will-formation theorized in this dissertation are not the whole story. But theorizing such *mostly* voluntary cooperation is a necessary concession to the reality of
human motivation, in which the rational integration of the acting self is an achievement and not guaranteed.  

The fact that the threat of punishment may, in moments of volitional instability, tip the balance toward public reasons does not mean, though, that coercion can replace rational commitment. That is to say, there is a difference between action that is primarily motivated by public reasons but needs a coercive “nudge” and action for which coercion creates a compelling (private) rationale on its own, such that public reasons are unnecessary. In the first case there is a significant source of motivation at work that simply isn’t there in the second. The lack of those commitments to public reasons may spell the difference between whether a given cooperative scheme will emerge or not.

Furthermore, some coercive sanctions can in effect, be an extension of the actor’s own will. The best-known examples of such sanctions are “commitment devices” by which actors attempt to bind their future selves to decisions made in the present. These measures straddle the border between voluntary and coerced cooperation. They are coercive in that at the moment the sanction is enforced, the actor is constrained against his will; but that constraint is itself the result of a previous voluntary choice.

Joseph Heath and Joel Anderson offer an interesting way to think about such devices. They may be considered part of a person’s “extended will”: a system of external aids or kludges

---

14 On this notion of integration, see Chapter 5 above. As Habermas argues, this is an important distinction between moral norms and legal norms (1996, 111–18). Through the “facticity” of their enforcement, legal rules, unlike moral rules, do not assume that compliance will be based only on moral insight and will. In Kant’s terms, law governs only the external aspects of action: it demands that people obey, and allows that they may be coerced if they do not, but does not require any particular motive for obedience. Kant 1991b.


16 Elster 2000. Such institutional commitment devices can be an important means of mitigating the commitment problems that undermine negotiation in violent conflicts. Walter 2009.
that she uses to augment her willpower.\textsuperscript{17} The idea is analogous to the “extended mind”—the idea that our minds should be understood to include material outside our heads that is part of our mental capacities, such as using a pencil to do math.\textsuperscript{18} In this sense, systems of incentives that actors deliberately create to help them remain resolute in the face of temptations to cheat could be considered part of the actor’s extended will.

For most actors in a “moral core, coercive periphery” system, coercive incentives will play a supporting role. But in stable cooperative institutions, there may also be actors who can be motivated only by coercion. These are actors who are not committed to public reasons at all: actors on whom public reasons have no motivational grip. From a sociological perspective, such actors would be considered “sociopaths”—though, as I argued in Chapter 6, in the context of an emerging political order, a failure to commit to public reasons is not necessarily the result of a failure of primary socialization. Political actors can be “normally” socialized actors who make and observe commitments within their own groups, while refusing to consider themselves similarly accountable to their opponents. Nevertheless, such actors may be brought to abide by shared rules and obligations to opponents by mobilizing coercive sanctions.

Again, we might ask whether these sanctions make normative motivations superfluous. Again, however, the answer is “no,” though this time for a different reason. At the level of a single actor who is utterly uncommitted to public reasons, cooperation will be motivated by coercion or not at all. But the same need not be true of the system of actors as a whole. That is because those who monitor and sanction cheaters are often themselves motivated to do so by a commitment to public reasons and shared rules. Monitoring and sanctioning cheaters tends to be

\textsuperscript{17} Heath and Anderson 2010.

\textsuperscript{18} Clark 1997; Heath 2014b.
subject to its own “second order” collective action problem.19 Among those who are committed to a system of rules, each benefits if cheaters are punished, though being the person who actually monitors or punishes is costly. It is not surprising, then, that cooperative schemes and systems of public reasons tend to include “second order” rules that prescribe the punishment of cheaters, as well as first order rules prescribing or prohibiting particular actions.

### 9.2.2 Backstopping cooperation

The second role of coercion in institutions for collective action is to *backstop* commitments to public reasons. Whereas “incentivizing” means that the actor is directly motivated by the possibility of being sanctioned, “backstopping” means that the sanction assures the actor that *others* will follow the rules. When coercion acts as an incentive, it gives her a private reason to comply; when it acts as a backstop, it supports her commitment to public reasons to comply.

Backstopping, in effect, enables people to “act on their own best impulses.” It can fill in for gaps in trust where actors are committed to public reasons, but are less than fully confident in others’ commitments—or, to use the terminology of Chapter 8, it can foster cooperation when *perception* lags behind *commitment*, and thus mitigate the problem of misplaced defensive defection.

The enforcement of rules is necessary to their normative validity. While there may be some moral norms that prescribe actions absolutely, regardless of whether others comply,20 that is not generally the form of the cooperative schemes that emerge between political opponents. If I follow the rules but someone else cheats, I have not only been harmed; I’ve also been wronged.

---


To be cheated on when there is no coercive recourse to see justice done is an affront, a moral harm as well as a material one. When someone cheats and gets away with it, the legitimacy of the system of rules is undermined.

The backstopping role of coercion is sometimes overlooked because it straddles the border between standard rational choice and sociological approaches to rules. The coercive backstopping of rules presumes the motivational effectiveness of both normative-rational commitment and coercive incentives, both public and private reasons. Those who are kept in line by coercive sanctions respond only to private reasons. Their actions are thus consistent with rational choice models of interaction and utility-maximization. But those whose commitments to public reasons are being backstopped or protected are genuinely motivated by those public reasons themselves. Thus, theoretical approaches that presume actors to all conform to one type or the other will overlook this important effect.

9.3 Coercion affecting will-formation: spillover and crowding out

The final set of effects to consider come about through interaction between the “moral core” and the “coercive periphery.” To this point in the chapter, I have implicitly assumed that voluntary and coercive modes of influence operate independently.\(^1\) In this section and the next, I will complicate that assumption by considering how coercive inducements affect voluntary commitments and how voluntary commitments affect coercive inducement.

When it comes to the effects of coercive inducement on voluntary cooperation, there are two kinds of effects to consider: (1) spillover, in which coercively induced cooperation

\(^{21}\) The exception was my brief mention of normatively motivated punishment, which is, in effect, a contribution of the moral core to the coercive periphery. I will have more to say about this in section 4 below.
contributes to voluntary cooperation based on relationships of mutual accountability; (2) the opposite, in which coercive sanctions *crowd out* and erode voluntary cooperation and relationships of mutual accountability (Figure 9.3).

Figure 9.3. Coercion affecting will-formation: spillover and crowding out

![Diagram showing the relationship between cooperation, spillover, and crowding out](image)

9.3.1 Spillover

The idea that coercively induced cooperation might spill over into voluntary cooperation is a venerable one. Kant placed this idea at the centre of his teleological conception of moral progress in human history. Kant argued that the “hidden plan of nature” is to use the devastation and sheer
exhaustion produced by civil war to lead people to establish civil societies and laws.\textsuperscript{22} With time, this “pathologically enforced social union is transformed into a moral whole.”\textsuperscript{23}

Spillover is also commonly seen as a mechanism of socialization.\textsuperscript{24} Here, norm-conformative behaviour that is initially induced through sanctions is eventually internalized, as the actor changes from experiencing the rule as an imperative imposed from the outside, to a norm that binds the actor \textit{in foro interno}.\textsuperscript{25} While the idea of internalized sanctions is used mainly (and perhaps most plausibly) as an account of the primary socialization of individuals, it has also been suggested as a mechanism by which states and armed groups can be socialized to respect international norms such as human rights.\textsuperscript{26}

How might such spillover occur? Returning to the categories introduced in the previous chapter, coercively induced cooperation could contribute to relationships of mutual accountability by fostering either the actors’ commitment to public reasons, or if they are committed, their perception that others are too. Both transformations are conceivable. They are perhaps most likely to occur in situations where sanctions are being applied reciprocally, by the actors involved to each other, rather than by a third party or outside intervener. In such cases, restraint imposed by each on the other may begin to look like, or actually become, self-restraint. When that happens, a pattern of behaviour that was merely a strategic equilibrium takes on a rudimentary normativity, so that each begins to see that restraint as something the other \textit{owes}

\textsuperscript{22} Kant 1991d, 50.
\textsuperscript{23} Kant 1991d, 44.
\textsuperscript{24} See Chapter 4.
\textsuperscript{25} Habermas 1987, vol. 2, chap. 1.
\textsuperscript{26} Risse and Sikkink 1999.
her, and as something she owes the other.\textsuperscript{27} When a strategic equilibrium comes to be seen not merely as the result of a struggle for power, but as being legitimate, then the relationship has begun to shift in the direction of mutual accountability.

A note of caution is in order, though. While spillover from induced to voluntary cooperation is possible, there is reason to expect such effects to be weak. The problem is that if political opponents know or suspect that each other is cooperating only because she is forced to, cooperation will not be normatively generative.\textsuperscript{28} When we know that cooperation is induced, we have no reason to ascribe good intentions to those who cooperate. As a result, if spillover occurs, it has to occur behind the actors’ backs—through mechanisms like habituation that operate beneath the actors’ conscious reasoning.

### 9.3.2 Crowding out

The opposite of spillover is crowding out.\textsuperscript{29} Here, coercively imposed sanctions \textit{undermine}, rather than contribute to, intrinsically motivated, reason-based cooperation. As a result, attempts to induce cooperation by offering incentives or punishments can backfire, resulting in less cooperation, not more.

There are a few ways crowding out can happen. First, the imposition of coercive sanctions can be taken as information that other people cannot be trusted: after all, that must be why it is necessary to threaten sanctions.\textsuperscript{30} In this way, crowding out stands in a paradoxical

\begin{footnotesize}
\begin{itemize}
\item\textsuperscript{27} Adler 2008; Kupchan 2010, 37–46.
\item\textsuperscript{28} Jervis 1976; Mansbridge 2009b; Luhmann 1979, 42.
\item\textsuperscript{29} Frey 1994; Frey and Jegen 2001.
\item\textsuperscript{30} Even talking about people as being generally self-interested can create this effect. Rothstein 2005b; Mansbridge 2009b.
\end{itemize}
\end{footnotesize}
relationship to the backstopping effects of sanctions I discussed above. In order for political opponents to feel that cooperation is safe, they must have some confidence that an adequate coercive response is available in the event of another’s cheating. But where sanctions become too salient in decision-making, they may shift from being a comforting assurance against unlikely cheating, which is consistent with the development of relations of mutual accountability and trust, to a sign that cheating is likely, which undermines potential trust.

Second, coercive sanctions can undermine intrinsic motivations by engendering resentment and disrespect. People who are intrinsically motivated by public reasons resent being treated as if they were not—and as a result, they may become less motivated to cooperate. Here, the imposition of sanctions not only undermines the perception that others are committed to public norms; it also undermines those commitments themselves.

The problem of crowding out suggests that the relationship between coercive capacity and cooperation may have a “sweet spot”: some coercion is needed to incentivize cooperation and backstop normative commitments, but too much coercion undermines those commitments and makes for less cooperation over all.

9.4 Will-formation affecting coercion: institution-building

Finally, we must consider the effects of commitment and voluntary action on coercive institutions (Figure 9.4). The primary such effect is institution-building. Institution-building is a joint project, something political opponents do together through deliberation and collective action. To this point in the chapter, I’ve examined the contributions of coercive sanctions to cooperation, taking their existence for granted. But of course, coercive sanctions themselves
have to come from somewhere. And in relatively anarchical contexts political opponents will have to build, or at least contribute to, those institutions themselves.

Figure 9.4. Will-formation affecting coercion: institution-building

Thus, it is not surprising that post-conflict institutions tend to be a central topic of negotiations in conflicts. In part this has to do with distributional struggle. Institutions often distribute the benefits of cooperation unevenly, and so their specific shape will be hotly contested. But the focus on institutions in negotiations also reflects the actors’ awareness that stable cooperation requires the support of coercive institutions. Coercive institutions play an important role in stabilizing cooperation, and since savvy political opponents know this, they
invest in building them. (And, as a corollary, where coercive institutions cannot be built—whether for technical reasons or otherwise—that can pose a significant barrier to cooperation.\(^\text{31}\))

Thus, the voluntary creation of new institutions to support collective action represents an important causal connection between voluntary and coercive interactions. In this way, coercive institutions can become part of a virtuous circle. As I argued in the previous chapter, voluntary cooperation begets further voluntary cooperation: as political opponents engage in deliberation and joint action, they simultaneously draw on and generate relationships of mutual accountability. But voluntary cooperation is related in a similar way to the generation of an institutional architecture that backs up voluntary commitments with sanctions. As opponents deliberate, work together, and increasingly see themselves as accountable to each other, they also build the relational resources to build and maintain institutions with coercive capacity. Those institutions, in turn, incentivize and backstop further cooperation, which can take the form of strengthening those institutions themselves. In this way, deliberative politics can feed into institutional capacities.

9.5 Conclusion

In this chapter, I have sought to explore the combinations of voluntary and coercively induced cooperation in stable institutions. While a deliberative conception of rule-based cooperation must focus primarily on collective will-formation and voluntary cooperation, it is important to acknowledge the contribution of coercion to stable cooperation as well. Due to the ever-present possibility of strategic defection in politics, and to the high stakes of being suckered, especially—

\(^{31}\) For example, Fearon (1998, 291–92) argues that the development of satellite monitoring in the early 1960s was, in part, what made it possible for the US and Soviet Union to negotiate the SALT I nuclear treaty.
in situations where violence is possible, schemes of cooperation that are based primarily on voluntary, reason-based cooperation still need to be supported by coercion.

Thus, as deliberative politics takes hold among a set of actors, coercion does not disappear, but rather becomes more legitimate and more productive. It is no longer the primary medium of politics. Instead, it plays a peripheral, supportive role. Coercion, now relatively legitimate to the extent that it is exercised in accordance with deliberatively achieved rules, remains as a check against the threat of deceit or violence, while the bulk of the work of coordinating social action is done through voluntary commitments, deliberation, and deliberatively sanctioned decision mechanisms. Coercive institutions backstop the relationships of mutual accountability that deliberative politics generates, and those relationships then allow people to cooperate in increasingly complicated, productive, and normatively legitimate ways.
Chapter 10  Conclusion: The Implications of Constituting a Public

How do shared rules and cooperative arrangements emerge in international politics, civil conflicts, weak states, and other unstable and relatively anarchical situations? It is hard to imagine a more pressing question for the people caught up in such difficult circumstances. In situations like these, especially where violence has taken place, the stakes of cooperation are undeniably high. The emergence of shared rules and mutual self-restraint often spells the difference between festering, intractable conflict—with its toll not only in lives, but also in livelihoods, well-being, and opportunities—and conflict that is tempered and channelled into a form of politics characterized by negotiation, compromise, and cooperation rather than unmitigated and destructive struggle.

In this dissertation, I’ve argued that deliberation is an important part of the answer to how shared rules and cooperative schemes emerge. In some respects, this conclusion may seem obvious. After all, conflict resolution and talk appear to go hand in hand. In assessing the prospects for cooperation in any difficult conflict, the first question we often ask is, “are the parties talking?” Of course, talking is no guarantee that an agreement will be reached, or once reached, upheld. Neither is all talk deliberative, as we have seen. Indeed, moments of genuine deliberation, understood as persuasion and compromise on the basis of reasons, may be relatively few and far between. But notwithstanding the difficulty of telling deliberative and non-deliberative talk apart, a considerable amount of the talk that surrounds the negotiation of

32 See Chapter 3.
cooperative rules appears, at least, to take the form of argument and more-or-less genuine attempts to persuade or convince.

But what is less clear is how deliberation might affect the emergence of cooperative rules, and this is where my dissertation offers a more surprising conclusion. We often take it for granted that if deliberation matters, it matters because of its effects on the content of the agreements reached. But that is not the only, and perhaps not even the most important, effect of deliberation. Deliberation can also contribute to the emergence of cooperative rules by helping to generate the shared commitments and mutual accountability that are necessary for political opponents to reach (and abide by) any agreement at all.

It is in highlighting and theorizing these often-overlooked relational effects of deliberation that my dissertation makes its main contribution. Deliberation is not just about the content of the reasons political opponents offer (or, from a normative perspective, should offer) to one another. It is also about the fact that they are offering reasons at all, and about what that implies for the kind of actions they can expect from each other. Most importantly, deliberation can be a way for interlocutors to establish and demonstrate their accountability to each other: their willingness to uphold, and count on others to uphold, shared standards of behaviour. Any such contribution deliberation makes is immensely important—especially in relatively anarchical situations where opponents’ basic commitments to abiding by shared rules are in doubt. In summary, the work that deliberation does is not only cognitive, in the sense of coming to legitimate agreements and shared understandings, but also relational, in that it establishes the social bonds, solidarity, and intersubjective commitments needed to act on those shared understandings.
The linchpin of this theory is a conception of the public and of public reasons. The idea of the public has long played a central role in liberal and democratic thought.  But while we are used to understanding publicity as a component of democratic legitimacy—or even as a test for the legitimacy of particular arguments, actions, or policies—we are generally less familiar with the implications of publicity for collective action. And yet publicity has profound implications for collective action, and for the ability of interdependent individuals to coordinate their actions on the basis of shared rules. As I have argued above, the existence of a public—understood as a group of actors who take themselves to be accountable to each other for the rightness of the reasons they act on—is, in large measure, what makes it rational, from each individual’s point of view, to abide by shared rules and agreements. To form a public is to see yourself as being accountable to others in such a way that what counts (to you) as a good reason for action comes to depend on what can be justified to others. The constitution of a public is thus the crucial link between shared reasons and shared actions, between deliberation and collective action.

10.1 The constitution of a public: wider implications

I emphasize this argument about publicity and collective action not only because it is the conceptual centre of my dissertation, but also because it has profound implications beyond my dissertation topic. The goal of the dissertation was to develop a deliberative theory of how cooperative rules emerge in relatively anarchical situations, but the resulting theory is highly

---


34 Kant 1991a, 130; Rawls 2005.

35 The core of that argument is contained in Chapters 5 and 6.
general in its applicability. My central argument about publicity and collective action should apply, in principle, to any collective action problem—which, as I argued in Chapter 2, are ubiquitous in political life. It amounts to a claim about one of the fundamental conditions for any rule- or norm-based cooperative scheme to work: namely, that any system of cooperation that depends upon individual normative commitments also depends upon those individuals seeing themselves as accountable to each other—which is to say, as forming a public. Moreover, the establishment of those relationships of mutual accountability is a contingent occurrence; it is something that must be achieved in any particular context of interaction. Thus, the constitution of a public is of fundamental importance in a wide range of political and social interactions.

The theory of collective action and publicity developed here makes a contribution to a broader argument that has been developing in political science and related disciplines since the discovery of the logic of collective action problems in the 1950s.\footnote{As Mansbridge (2001; 2010) points out, political thinkers since Plato and Aristotle have recognized the challenge that self-interest can pose to realizing the common good. What is revolutionary about the modern, game-theoretical formulation is that it recognizes this not (or not only) as a problem of individual motivation (i.e. individuals are insufficiently virtuous or public-spirited) but as a problem generated by the structure of certain \textit{goods}, namely, goods that are non-excludable.} That argument begins with the recognition of collective action problems as basic, widespread, and serious. As Jane Mansbridge argues,

whenever people must deal with a \textit{non-excludable} good—a good from which non-contributors cannot be excluded—there is an inbuilt trajectory to failure in collective action. This is probably the most important problem of human organization. It is certainly the most important problem of our time. It shapes the dynamics of global warming, the environment generally, nuclear proliferation,
taxes, health care provision, and most other issues for which human beings have
devised (or failed to devise) public governing systems.\textsuperscript{37}

The second part of this larger argument is the more recent consensus, emerging among a
growing body of researchers across the social sciences and humanities, that \textit{normative}
\textit{constraints} play a central role in overcoming collective action problems.\textsuperscript{38} This development
represents a decisive shift away from “classic” rational choice assumptions, in which choices
were assumed to be explicable in purely self-interested terms.\textsuperscript{39} Some of the more compelling
evidence for the importance of social norms has come from laboratory studies and the study of
human evolution. In laboratory settings of collective action problems, people consistently
cooperate at levels far above what is predicted in the standard “self-interest” models, and they do
so, moreover, in ways that indicate that they are responding to pro-social norms.\textsuperscript{40} In the study of
human evolution, there is growing evidence that the “ultra-sociality” of human beings—that is,
our ability (apparently unique among animals) to sustain very large-scale social cooperation
among genetically unrelated individuals—is founded on a disposition to conform to social

\textsuperscript{37} Mansbridge 2010, 590. See also Mansbridge’s (2012) Madison Lecture to the American Political Science
Association.

\textsuperscript{38} Bicchieri 2006; Bicchieri, Duffy, and Tolle 2004; Fehr and Fischbacher 2004b; Fehr and Gintis 2007; Frohlich
Ostrom 2005; Ostrom 2009. This shift also has an analogue in the rise of constructivist IR theory and its criticism of
the rational-choice-inspired neo-realist and neo-liberal theories that preceded it.

\textsuperscript{39} More precisely, the truly decisive shift occurs with the idea that people act on \textit{deontological} normative
considerations. Consequentialist normative considerations, such as altruism, can be integrated more easily into
classic rational-choice models by simply adding to the “utility” actors derive from outcomes that benefit others. By
contrast, deontological normative values are attached not to outcomes but directly to actions themselves, which
creates a structure of decision-making that breaks more fundamentally with classic rational choice models. Heath
2008.

\textsuperscript{40} Ostrom 2000; Heath 2008, chap. 6.
norms, and, moreover, is not likely the result of the sort of rational “tit-for-tat” strategies theorized in the literature on repeated games.41

While I cannot offer a defense of this growing consensus here, the theory of publicity and collective action developed in this dissertation is consistent with this line of argument. That is simply to say that I offer an account of collective action that takes norms, and normative commitments, seriously. But my theory of publicity also adds an important third step in this broader argument. That third step is to recognize that in any particular context of interaction, the force of those social norms is itself a contingency that depends on the constitution of a public. If people take themselves to be members of a common moral community, a public, the public reasons of that community will have a motivational grip on them, and those shared normative standards often play a decisive role in maintaining cooperation. But all that depends on the actors involved seeing themselves as accountable to each other in the first place. Such moral communities have to be formed, and they do not always exist where, from the point of view of collective action, they should. As I argued in Chapter 4, in many political conflicts, the basic normative commitments and social bonds that underpin the enterprise of making and following rules together cannot be assumed; rather, they must be built.

The constitution of a public, then, is a very basic phenomenon in political life. Since the discovery of the logic of collective action in the 1950s, many observers of politics now ask, when approaching a new issue or conflict, “is there a collective action problem here?” If the argument of this dissertation is correct, the very next question they should ask is, “if so, is there a public here?” because the answer to that question does much to determine the prospects for cooperation and the kind of politics that is likely to emerge. To the extent that the people

involved do see themselves as a public, it will be possible for them to both speak and act together on the basis of public reasons. To the extent that they do not, they will be restricted to forms of purely instrumental interaction and communication, and thus to the kind of tragic results that game-theoretical models predict. In that case, productive politics depends in large measure on constructing a public, and thus on the processes I have begun to theorize in the chapters above.

10.2 Some final thoughts on democratic theory

I want to conclude by pointing to some of the democratic implications of this work. The study of deliberation in political science is closely associated with democracy, and yet democratic concerns have remained very much in the background of my work here. Rather, the deliberative theory developed above is framed as a response to collective action problems and to the challenge of building cooperative institutions in relatively anarchical circumstances.

Collective action, however, is itself essential to democracy. Unmitigated collective action problems are not only harmful to people’s interests, they also create democratic deficits. A fundamental democratic ideal is that people should not be subjected to the arbitrary power of others. And yet where people are subjected to interdependencies they do not choose, that is precisely the effect. Democracy requires giving people the ability to contest the harmful effects of others’ choices—and not only to protest and resist, but also to act together and respond to problematic interdependencies collectively. In recent years, democratic theorists have thus

------------------------

42 Chapter 6.
43 Pettit 1997; Bohman 2007; Lovett 2010.
44 Mansbridge 2012; Warren 2012b.
become increasingly attuned to the ways in which democratic action and collective action are connected.

Part of the recent interest in collective action problems among democratic theorists is a response to what have been called “the new circumstances of politics,” meaning the increasing complexity and global connectedness of modern societies. Complexity refers not only to technical complexity but also to fact that decision-making in modern societies is increasingly dispersed away from centralized state institutions and into myriad economic and social organizations, from semi-independent regulatory bodies to industrial associations, unions, corporations, hospitals, and universities. Added to that is the growing importance of forms of interdependency that cross national borders, from global supply chains to financial instability and climate change. In light of these trends, theorists have increasingly questioned the adequacy of thinking of centralized and territorially bound state institutions as the primary locus of democratic action. The result has been a shift “from demos to demoi,” that is, from the model of a single democratic people legislating for itself to a more fluid conception of multiple and overlapping peoples emerging in response to particular collective action problems. On this view, democratic citizens may, in effect, belong to any number of “peoples,” each of whom is composed of those individuals whose lives are affected by each other in a particular way, often across state boundaries.

47 Warren 2010; Fung 2013.
48 M. S. Williams 2009.
The conception of publicity developed in this dissertation offers some important conceptual resources for this emerging project in democratic theory. In particular, it offers a way of thinking about how such interdependent individuals might come to see themselves as democratic peoples, that is, as groups capable of acting together and making collective decisions. The problem for a democratic theory focused on demoi defined in terms of “communities of shared fate” or affectedness is that being affected by a common collective action problem does not, of course, guarantee that people will be willing or able to respond to it. The theory developed here suggests one way of theorizing how such an acting “people” is formed. What makes groups of interdependent individuals into democratic peoples may be the same thing that allows political opponents to act together in anarchical circumstances, namely, the emergence of relationships of mutual accountability among them—which is to say, the construction of a public. While the notion of publicity has long had a central place in liberal-democratic thought, it tends to be used primarily as a normative standard by which to judge the legitimacy of particular arguments, institutions, or policies. That is a legitimate and important use of the concept. But the idea of publicity also holds great promise as a way of understanding how interdependent individuals come to form and maintain the democratic publics or peoples that can act together in the first place.

Bibliography


The Economist. 2011. “A Plaything of Powerful Nations.” Published October 1, p. 64.


