THE MORAL SIGNIFICANCE OF SOCIAL CONVENTIONS

by

CHRISTOPHER JOHN MACDONALD

B.A. (Hons.), Trent University, 1992
M.A., The University of British Columbia, 1994

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Department of Philosophy)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

February 1999

© Christopher John MacDonald, 1999
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy

The University of British Columbia
Vancouver, Canada

Date April 29, 1999
Abstract:

This Thesis is about the role which social conventions play in shaping our moral choices, and about the possibility of a normative theory that takes such conventions seriously. It also hints at the idea of looking at conventions as a kind of moral technology. If existing conventions serve a useful function, then perhaps we can take what we know about conventions and apply that knowledge in a forward-looking manner.

Chapter 1 of this Thesis outlines the shape of the project, and explains its roots in methodological individualism and a relative, subjective theory of value. Chapter 2 surveys the literature on norms and conventions, and explains why it is that despite the prevalence of the former in the literature, moral theorists should focus on the latter. Chapter 3 looks at the ways in which conventions structure strategic interaction. They do so both by providing information that served as an input to rational decision making, and by providing a more direct, non-instrumental form of motivation. In Chapter 4, we look at the relevance of the literature on public goods and the problem of collective action, and argue that beginning and maintaining conventions embodies a collective action problem. In Chapter 5, we move to examine the normative force of conventions, and suggest that conventions constitute a reasonable set of constraints on self-interested behaviour. Chapter 6 addresses the fact that a convention-based approach seems capable of endorsing asymmetrical outcomes, at least some of which seem unjust. In Chapter 7, we look more generally at the adequacy of a conventionalist account, and examine the departures which such an account seems to make from everyday morality. Finally, in Chapter 8, we put theory to practice, and examine the world of professional obligation. We offer there a meta-ethics of professional obligation that suggests that such obligations are best understood as conventions between professionals. That argument serves as a challenge to existing principle-based and virtue-based theories of professional obligation, and illustrates the practical importance of this Thesis. In Chapter 9, we summarize the ground covered, and discuss how we can apply what we know about conventions to ameliorate problematic situations.
**Table of Contents:**

Abstract .................................................. ii
Table of Contents ........................................ iii
List of Figures ............................................. iv
Acknowledgements ........................................ v
Chapter 1: Introduction ................................... 1
Chapter 2: What are Soft Institutions? ................. 23
Chapter 3: How Conventions Affect Choices: Knowledge and Motivation ........................................ 47
Chapter 4: Conventions, Public Goods and Collective Action .................................................. 73
Chapter 5: The Moral Significance of Social Conventions .................................................. 87
Chapter 6: Beyond Symmetry ................................ 114
Chapter 7: The Limited Aspirations of a Convention-Based Account ........................................ 137
Chapter 8: Clinical Standards and the Structure of Professional Obligation .................................. 161
Chapter 9: Conclusion ........................................ 173
Appendix 1: Games .......................................... 182
Bibliography .................................................. 185
| Figure 3.1: “Bank Customer’s Dilemma” | 57 |
| Figure 3.2: “Bank Customer’s situation, given broad willingness to impose sanctions” | 61 |
| Figure 5.1: “Bank Customer’s Dilemma” | 93 |
| Figure 5.2: “The Strategic Structure of a Convention” | 94 |
| Figure 6.1: “Asymmetrical Trade Relationship” | 118 |
| Figure 6.2: “Moral behaviour in an asymmetrical trading relationship.” | 119 |
| Figure 6.3: “The Battle of the Linguistic Groups” | 125 |
| Figure 7.1: “The Moralist’s Dilemma” | 158 |
| Figure 8.1: “The Doctor’s Dilemma” | 167 |
| Figure 8.2: “Contract Between Society and Physicians” | 171 |
Acknowledgments

I will begin by thanking my Committee, who collectively provided absolutely optimal supervision. Professors Peter Danielson, Paul Russell, and Wayne Norman gave me the guidance I needed while at the same time respecting my abilities enough to give me a very free rein.

The greatest single debt I owe is to my supervisor, Dr. Peter Danielson. Peter’s work has had an enormous influence on my thinking over the time I have spent studying with him. It would be impossible to give due credit for every one of Peter’s ideas that made it into this Thesis, although I have tried to give specific credit for a few of his more noteworthy and original suggestions. Peter’s generous comments on drafts improved the final product in innumerable ways. I thank Peter for being such a keen critic, for setting the bar high, and for thinking highly enough of my work to consider it worthy of his best criticism.

I would also like to thank several organizations for their financial support during the researching and writing of this Thesis. I benefited greatly from the combined support of: a University Graduate Fellowship from UBC; a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada; a Teaching Assistantship from the UBC Philosophy Department; and a Research Assistantship from the UBC Centre for Applied Ethics. My inclusion of the Centre for Applied Ethics in this list threatens to underplay the importance that organization has had for my graduate education. The completion of this thesis, along with my general academic and professional development, owes much to both the generous financial support and the unending intellectual stimulation provided by the Centre. I owe the Centre, and particularly its Director, Dr. Michael McDonald, and enormous debt of gratitude.

I presented sections of this thesis to several audiences. A version of Chapter 4 was presented to the UBC Philosophy Department Undergraduate Society’s Socratic Sphere. Versions of Chapter 5 were presented to UBC Philosophy Department’s Graduate Colloquium and the Western Canadian Philosophical Association. An early draft of Chapter 8 was presented to the UBC Centre for Applied Ethics Colloquium Series; a later draft was presented to the Canadian Bioethics Society. I would like to thank the members of all of these audiences for their stimulating and encouraging comments.

My office-mate, Bryn Williams-Jones, generously proof-read versions of every chapter, and offered a number of helpful suggestions. William Harms’s comments on what eventually became Chapter 5 were instrumental in helping me clarify the nature of this project.

Peter Lee and his staff at The Well coffee shop on Dunbar Street in Vancouver have my undying thanks for providing a seemingly endless supply of good coffee and good cheer.

Finally, I dedicate this work to Kimberly Proulx, who provided invaluable support in too many ways to mention.
Chapter 1: Introduction
“The practice of the world goes farther in teaching us the degrees of our duty, than the most subtle philosophy, which was ever yet invented.”
David Hume, *A Treatise of Human Nature*, Book 3, Part II, Sec. XI (pg. 569)

Scenario 1:
Mr. Beta’s alarm went off at sunset. Having slept all day, he arose in preparation for a night’s work at his private engineering consulting practice. As he drove down Highway 417 into Ottawa, his speed varied between 80 kph (a speed that provoked a great honking of horns on the part of his fellow motorists) and 150 kph (a speed that induced a different sort of anxiety on the part of his fellow drivers). Today he chose to drive on the right side of the road, though yesterday he had chosen differently. He arrived at the office three cigarettes later – just in time to greet his first client of the evening. He offered his left hand just as the client thrust forward his right. He then offered his right hand, just as the client switched to offer his left. As both settled for simply nodding their hellos, the client wondered whatever possessed Mr. Beta to do his business at night. When asked, Mr. Beta replied that he didn’t see what difference it made, though he admitted that for some reason he had trouble filling his midnight-to-six appointment slots, and that clients tended to react badly when he called them during those hours. After work, he headed to a favourite greasy-spoon for a bite to eat. He had eaten here with friends before, but they weren’t there this time. His meal cost $4.50. He handed the waitress a five and waited for his change, which he pocketed as he left. Mr. Beta had told a friend that he’d meet her for breakfast later this morning, but having already eaten he decided to go home instead.

Scenario 2:
Mr. Alpha’s alarm went off at 7:00 a.m. Having been asleep all night, he arose in preparation for a day’s work at his private consulting practice. He drove down Highway 417 into downtown Ottawa, at a speed of exactly 110 kph (10 kph over the limit – no more, no less). He drove, of course, on the right side of the road. He arrived at the office at 8:15 – just in time to greet his first client with a hearty right-handed handshake. After work, Mr. Alpha met some friends at the local pub; it was Wednesday, after all. And since it was the second Wednesday of the month, Mr. Alpha proceeded to order and pay for the evening’s first round of drinks (Janet had bought the first round last Wednesday, and Carl would buy it next Wednesday). He made sure to give the waitress a fair tip. Another friend, not part of the usual Wednesday night crowd, showed up soon after, since Mr. Alpha had assured him the day before that he’d be in this pub at this time. A good time was had by all.

By now, the reader may well have guessed the point of these two narratives. Mr. Alpha’s day was the sort of day lived by millions of people on a regular basis: a day guided
by standards or norms. Almost every sentence of the description of Mr. Alpha’s day involves some adherence to one standard or another. And almost every sentence of the description of Mr. Beta’s day involves some disregard for standard ways of doing things. In fact, each of the behaviours described above is a matter of convention, both in the everyday sense and in a more technical sense to be described below. The 12-hour clock; our generally diurnal work habits; shaking hands as a form of greeting; shaking with our right hands instead of left; driving on the right side of the road; local variance from posted speed limits. All of these behaviours will readily be admitted to be conventions in the usual sense; that is, we do them as something like a matter of collective habit, and individually we do them because everybody else does them. Others of these behaviours – those involving compensatory justice and keeping promises, for example – may be less obviously conventional, since they move us from the domain of the mundane into that of the moral. But I will argue in this thesis that these, along with many other apparently morally required actions, are conventional in nature.

Traditionally, moral philosophers have looked at difficult situations, and asked how we can think our way out of them. This project begins instead by looking at ‘easy’ situations, asking what conventions make them easy, and asking how we can replicate that ease by building new conventions. The main arguments of this thesis will be to the effect that conventions structure our social interactions and delimit our moral obligations.

The term ‘convention’ has been much abused in common parlance and in various literatures. It is often used as a synonym for ‘custom’ or ‘fashion’ or ‘habit.’ As used by

1 The reader may also have noticed that Mr. Beta is named after a technology – Betamax video – that failed to achieve the status of a convention.
moral philosophers, it most often has a derogatory flavour; the principles justified within the framework of one or another moral theory, for example, are often contrasted with "mere social convention."² Both the sloppy usage and the pejorative usage are unfortunate. This thesis seeks to reclaim this undervalued concept, for use in both moral theory and applied ethics. For our purposes here, we will give the concept more content, differentiating conventions from traditions, customs, norms, prejudices, and other such regularities of behaviour and belief. Conventions, here, will be defined in the Humean sense as regularities of behaviour, sustained by a shared interest in coordination and an expectation that others will cooperate.³ Conventions, then, can be defined in terms of these three key concepts: regularity of behaviour; an interest in coordination; and an expectation of cooperation. The degree to which various conventions will embody these three characteristics will vary, but all three characteristics will always be present to some degree. Examples include the everyday conventions of lining up at banks, keeping promises, and buying alternating rounds of beer at the pub, as well as such serious conventions as avoiding killing non-combatants in war. Each of these ways of acting is useful for all concerned; none of them is strictly arbitrary, nor is any of them necessarily optimal. Like these examples, many of our accepted moral obligations cannot be understood except as the product of a convention. They are best seen as adequate responses to social exigencies. The unifying characteristic of each of these situations is, roughly, a surprising degree of social

² Further, many historically important moral philosophers (e.g., Bentham, Mill) were also social reformers, and denigration of convention would suit their rhetorical purposes and political goals.

³ I will leave the task of a more formal definition until Chapter 5. As Hardin (1995, p.113) writes, "[t]o impose a definition first and then look for explanations is generally backwards. Ostensive definitions of the form 'that is a norm, and that, and that,' are adequate for many of our categories until we understand them much better than we do."
cohesion or order. In each situation, the interests of the individuals involved would, on some description of the situation, lead us to expect a very different – much more turbulent – outcome.

Note also that implicit in the definition of a convention is the fact that any given convention will be but one of several possible ones for that situation. That is, it is not a regularity of behaviour strongly determined either by human nature or the environment. But neither (contrary to one common usage of the term) are conventions entirely arbitrary. Some norms may be arbitrary, but conventions never are. As defined here, a convention is sustained in part by an interest in coordination: thus conventions always serve some useful purpose. Gauthier agrees with this strengthening of the term. Gauthier (1977, p.333-4) writes that “...to suppose that what is conventional must therefore be arbitrary is entirely contrary to the spirit of contractarianism, which finds only in convention a sufficient rationale for society.” Further, “[t]o consider society arbitrary is to suppose that it affords no sufficient fulfilment or meets no fundamental need of most or all of its members.”

It has been argued that the primary function of morality is to overcome some major obstacles to getting along in social situations. The first obstacle has to do with limitations of human sympathies — that is, not only do we tend to be selfish, we tend to be selfish in short-sighted ways that get us into trouble. The second obstacle has to do with certain problems of collective action; even well-intentioned individuals cannot always achieve desired outcomes when those outcomes require cooperation on a large scale. We can identify two main sorts of solutions that have historically been offered to these problems. On the institutional side, Hobbes held that these problems could only be overcome by means of the coercive institutional apparatus of the state. On the individualistic side, some
have argued that what is needed instead are good dispositions or moral virtues — good individuals need no institutional mechanisms to help them get along. Others, like Kant, have held that individual rationality would be sufficient to dictate appropriate behaviour to avoid the problems inherent in social living. In a similarly individualistic vein, some thinkers (drawing on Adam Smith) have thought that the invisible hand of the market would be sufficient to turn individual self-interest into collective order.

Let us first briefly review some worries associated with Hobbesian institutional solutions. The first worry is a pragmatic one: few problems are amenable to direct government intervention. The first half of this worry is epistemic. It is unlikely that any central institution will have sufficient information gathering capacities to guide extensive regulatory efforts. The ability of governments to solve social problems is also limited by a lack of understanding of which policies — or even which policy tools — are best suited to solve particular problems. The second half of this worry has to do with limits on the effectiveness of government intervention in influencing individual behaviour. Government cannot be everywhere at once.

The second worry with the Hobbesian solution is, of course, the moral one. Although most would agree that, at some level, the law and order provided by government is a necessary precondition of personal autonomy, most would similarly agree that beyond some minimum, additional government intervention decreases rather than increases autonomy. Even if government could be everywhere, we would not want it to be.4

Next, let us review the problems associated with individualistic solutions. Do Kantian appeals to individual rationality work to solve these problems? The most general

4 For an excellent examination of the justifiability of government, see (Schmidtz 1991).
argument against this sort of solution is that rationality often underdetermines choice; for a large range of important situations, we need to know how others will act in order to decide what to do. The second sort of individualistic solution, namely Aristotelian appeals to individual virtue, must fall prey to our general pessimism about the prospects for altering human dispositions on a large scale.⁵

Next, let us look at our third class of individualistic solutions, namely the market-based solutions of Adam Smith and his followers. Smith held that order could come out of chaos through the actions of many individuals each pursuing his or her own good. Though this view has been shown to have considerable merit, faith in the market's power to create order is limited by social dilemmas such as the Prisoner's Dilemma, not to mention an understanding of the logic of collective action, which suggests that even well-intentioned individuals can fail to meet their objectives when doing so requires the cooperation of a large number of people. These forces together imply the likelihood of market failure. Further, it has been argued that the market, rather than creating order, is itself dependent upon a pre-existing system of property rights sustained either by convention or by law.

So we see that proposed solutions to the problems of social living have focused on the largest and the smallest units of analysis, namely governments and individuals. Yet it seems clear that a significant degree of order is generated at a variety of levels between these two extremes. Social order can be found in the wide range of micro- or meso-institutions that abound in modern societies: businesses, churches, social clubs, clans or

⁵ As Hume (Treatise, III.IV, p.521) put it, if politicians and moralists seek to change human behaviour by correcting natural human selfishness and ingratitude, "...they wou'd never make any progress, unless aided by omnipotence, which is alone able to new-mould the human mind, and change its character in such fundamental articles."
families (however defined), community groups and other political interest groups. But social order can be generated by a number of less formal structures, including norms, customs, traditions, fads, and habits. I will refer to these less formal structures as soft institutions.6

My project lies in between the institutional and individualistic paths, and involves seeking good conventions — one species of soft institution — to help us get along. This approach should be attractive, given on the one hand general worries about autonomy that arise whenever we contemplate relying too heavily on the coercive power of the state, and given on the other hand general pessimism about the prospects for altering human dispositions on a large scale. Conventions lie in between these extremes, having the advantages of being less psychologically demanding and more certain than reliance upon individual restraint, and less coercive than government regulation.

Since my focus is on institutions — social structures — that affect everyday moral decisions, my project lies somewhere between traditional moral theory and political theory. Not surprisingly, it draws on scholarship from a range of disciplines. It draws for example on work by philosophers such as David Gauthier, Joseph Heath, and George Mavrodes; political scientists such as Russell Hardin and Jon Elster; and economists such as Ken Binmore, Arthur Denzau and Douglass North, and David Hirschleifer. The historical roots of the project are in the writings of Thomas Hobbes, Adam Smith, and especially David Hume.

Among existing streams of modern moral thought, social contract theory is the one to which my project is most naturally compared. In a way, this project implies a critique of

---

6 I discuss — and justify — the notion of a ‘soft’ institution in Chapter 2.
social contract theory: the definition of a convention says nothing of equality or of rational agreement. However this critique is generally a friendly one: I see social contract theory (among the leading moral theories) as providing the most plausible account of our moral obligations. Social contract theory is of course far from monolithic; however, its various strains share certain strengths and weaknesses.

What (Hobbesian) contractarians get right is the idea that morality is an artefact, and that it is "justified only to the extent that it effectively furthers human interests" (Hampton 1995, p.159). As Gauthier (1991, p.18) argues, we no longer accept the world-view on which Kantian/Christian conceptions of morality depend. Contractarians generally err, however, to the extent that they fail to see the importance of considering moral agents as historically and socially situated. Social contract theory mistakenly assumes, for the most part, that agents (or societies) are in a position of radical choice vis-à-vis their moral principles, or more generally, their social arrangements. Contractarians, in their focus on rational choice, have also overestimated the importance of certain kinds of psychologies, and have thus neglected the importance of external factors such as established social conventions. Robert Sugden argues that social contract theory and conventionalism go hand in hand. According to Sugden (1990, p.769), "If contractarian reasoning can generate any moral conclusion at all, it must be that morality is, in important respects, a matter of convention."

Perhaps the best way to illustrate the main points of this Thesis is by means of an example. One of the most striking examples is provided by Mavrodes’s (1990) examination of the prohibition on killing non-combatants (e.g., civilians, medics, chaplains, and
P.O.W.s) during times of war. Mavrodes points to this widely acknowledged special moral status of non-combatants, and argues that most proposed grounds for this special status fail upon rigorous philosophical examination. Mavrodes (p.119) notes that, according to theorists who defend this special moral status, “the moral immunity of noncombatants consists...in the fact that their death can never, morally, be made the intended consequence of a military operation.” This immunity cannot be cast in terms of a general prohibition on killing. Most will agree that killing is sometimes justified; the thesis to be examined here is rather that this particular group cannot justifiably be intentionally killed. As Mavrodes (p.120) observes, it seems that in order to justify this special status, we must find some morally relevant difference between combatants and non-combatants. Such a distinction, according to Mavrodes, is surprisingly hard to find.

The theorists with whom Mavrodes is concerned (and perhaps most non-theorists) have cast this distinction in terms of guilt and innocence, and of a duty not to kill innocent persons. But as Mavrodes (p.123) points out, any reasonable account of what it means to be innocent fails to accord with the distinction between combatants and non-combatants: there will often be armed conscripts who “have no understanding of what the war is about, and no heart for it,” just as there will often be civilians – from farmer to stockbroker – who ardently support (and even hope to gain from) the war effort. Non-combatants often

---

7 There is room for doubt as to the pervasiveness of this prohibition. As Mavrodes (p.128) notes, for example, it was to some extent neglected during World War II (not least by the British and Americans in their bombing strategies). But the strength of this prohibition is evident in the fact that departures from it, when they become public, still seem to require special justification.

8 Mavrodes (p.119) notes that most of the theorists to whom he refers hold that the unintended killing of noncombatants (what has lately been called “collateral damage”) is not utterly forbidden.

9 Mavrodes’s examination of principled justification for this special moral status stops here with the failure of deontological reasoning, though he does not specifically identify this reasoning as deontological.
contribute to war efforts in a variety of ways, yet for the most part they are not subject to the lethal force applied to those who carry guns to the front lines.\footnote{And as Ignatieff (1998, p.24) notes, “in modern civil wars – Lebanon in the 1980s, Bosnia and Rwanda in the 1990s – where the distinction between civilian and combatant is often blurred and neighbor kills neighbor, it is difficult to distinguish the innocent and the guilty.”}

We can add to Mavrodes’s assessment by pointing out that the immunity of non-combatants cannot readily be explained on utilitarian grounds, any more easily than it can on deontological grounds. For on utilitarian grounds, one would seem to be justified in killing non-combatants in any case where that seemed likely to promote the best outcome overall; it is easy enough to imagine plausible circumstances where this would be the case. Indeed, it is a standard criticism of utilitarianism that it utterly fails to protect the innocent from being used as means to various ends. Thus even if we were able to parse more finely the moral responsibility that each citizen has for a particular war, utilitarianism would not be able to justify a firm prohibition on killing even the truly innocent. Would moving to rule utilitarianism help, here? This is perhaps more plausible, but it is entirely unclear whether a rule mandating the immunity of non-combatants would in fact be utility-maximizing. Indeed, it might seem rather that it would be utility-maximizing to have a rule mandating all-out, no-holds-barred war, so that war became such a terrible prospect that no country would risk it.\footnote{Indeed, Ignatieff (1998, p.116) hypothesizes that in pre-WWI Europe, the belief that various recent agreements on the conduct of war would ‘civilize’ war may in fact have made war more likely.}

Finally, it seems clear that the immunity of non-combatants cannot be accounted for in terms of social contract theory either. In order for social contract theory to generate an obligation, there must be (ideally) agreement or (minimally) some history such that the
parties involved can be seen as having made an implicit agreement. Certainly the immunity of non-combatants is taken to cover cases where there is no agreement – indeed, no history of direct interaction at all – between two countries. Further, the sorts of agents involved here are not the sort readily covered by social contract theory; that is, they are nations rather than individuals.

So it seems that the immunity of non-combatants cannot readily be reduced to the sorts of reasons endorsed by the major moral theories of our time. Thus we are left, with Mavrodes, to suggest that this special moral status is in fact a matter of convention. Mavrodes arrives at this conclusion by means of a thought experiment. He asks us (p. 124) to imagine “a statesman reflecting on the costliness [in human terms] of war.” Mavrodes suggests that limited combat (i.e., combat in which non-combatants are spared) is one of a range of conflict-resolution conventions that might arise that would be on one hand less costly in terms of human life than all-out war, but on the other hand more decisive (i.e., more effective at reducing capacity for further conflict) than combat of single champions.

Thus it is not that this special moral status is conventional in the everyday sense of being strongly arbitrary. Rather, it is conventional in a sense that, as Mavrodes (p. 126) puts it,

“a given obligation is convention-dependant if and only if (1) given that a certain convention, law, custom, etc., is actually in force one really does have an obligation to act in conformity with that convention, and (2) there is an alternative law, custom, etc. (or lack thereof) such that if that had been in force one would not have had the former obligation.”

12 There is, of course, another logical possibility: there may in fact be some principled grounds for this special moral status that is simply not captured by existing moral theories. Clearly the burden of proof lies with those who would make such a claim. Given the flexibility of the major moral theories, I take their collective failure as damning.
That is, the obligation to respect the immunity of non-combatants finds its source – in some sense or another – in the widespread practice of respecting the immunity of non-combatants. As Mavrodes (p.128) puts it, “[i]n cases of convention-dependent obligations the question of what convention is actually in force is one of considerable moral import.”

For some matters, at least, the absence of a relevant convention seems to imply the absence of obligation. In particular, we see this in the crucial case of a nation faced with waging a just war against a neighbour who is unwilling to exercise any constraint vis-à-vis non-combatants. In such a situation, it is hard to see why one would be morally obligated to grant the neighbouring country’s non-combatants any special status. Of course, a nation in this position might find it politically expedient to show constraint, in order to preserve a certain image of itself on the international scene. Less cynically, the leadership of a nation might exercise restraint because the citizenry demanded that their nation take the moral high road. The citizenry might insist that their soldiers not lower themselves to the level of their morally bankrupt enemies. The leadership thus might feel morally obliged (as good democratic representatives) to meet the reasonable demands of the populace. This, of course, would explain why the leadership of the nation had an obligation to grant the enemy’s non-combatants special status, but it would not explain why the citizens thought they had an obligation to demand this of their leaders.

13 With Mavrodes (p. 117) I will “assume without discussion that under some circumstances and for some ends warfare is morally justified.”

14 I am indebted to Bryn Williams-Jones for pointing out this possibility to me. His excellent example was Operation Desert Storm: American TV audiences were shocked at depictions of the infliction of “collateral damage” on Iraqi civilians, despite Saddam Hussein’s own disrespect for the lives of Kurdish civilians.
Hume, too, held that limitations on the conduct of warfare are a matter of convention. Indeed, in Book III, Part II, Section XI of the Treatise, Hume offers the “laws of nations” more generally (i.e., the laws governing the interaction of nations) as evidence of the conventional nature of justice. Hume points out that interaction among nations, like interaction among individuals, necessitates the observance of certain mutually advantageous rules. But according to Hume, it is widely acknowledged that the obligations of one prince to another are much less strict than those governing the actions of private individuals. Hume attributes this difference in the strictness of moral obligation at the two levels to a similar difference in the need for rules, noting (Treatise, III.II.XI, p.569) that “tho’ the intercourse of different states be advantageous, and even sometimes necessary, yet it is not so necessary nor advantageous as that among individuals, without which ‘tis utterly impossible for human nature ever to subsist.” The fact that the strength of obligation seems proportional to the degree of necessity implies for Hume that the obligation has its source in necessity – that is, that obligations are social artefacts that arise out of a need for mutual constraint.

This discussion of the conventions of war invites a brief mention of that most famous convention of all, the Geneva Convention. According to Ignatieff (1998, p.112) the twelve nations who signed the 1864 Geneva Convention, “agreed to ‘neutralize’ hospitals, ambulances, and medical staff, and...established the principle that enemy

15 Hume’s examples (Treatise III.II.XI, p.567) are “the sacredness of the persons of ambassadors...[and]...the abstaining from poison’d arms....”
16 Hume’s other example in this regard concerns duties of modesty and chastity. In a prescient bit of sociobiology, Hume argues that women are regarded as having stricter duties of modesty and chastity than men due to the social need to assure men of the paternity of their children. See Treatise III.II.XII.
17 All details of the Convention are taken from Ignatieff (1998).
soldiers deserved the same medical treatment as troops of one’s own nation” thereby setting “a standard that combatants had to meet if they wished to be thought ‘civilized.’”18 The Convention, though a formal signed document, was also a convention in the less formal sense discussed here: according to Ignatieff (1998, p.112) the Geneva Convention “did not fix any penalties for noncompliance, and it had no mechanism for enforcement.”19 Here we see the significance of the etymology of the word ‘convention.’ The word comes from the Latin “convenire,” meaning “to come together” and one of its uses is to describe a group of people physically meeting in one place.

**The Method:**

The general method employed in this Thesis is constructive, rather than critical. This constructive method reflects in part the relative dearth of work on the ethical importance of social conventions: critique is hollow where there is not really even a cluster of well-established views, let alone a standard doctrine. This approach also reflects a personal – professional – methodological preference. Good criticism is important, but ultimately insufficient. The goal of this Thesis is rather to sketch the details of a moral

---

18 The 1864 Convention was updated in 1949 by the establishment of four separate treaties, known collectively as the Geneva Conventions. These conventions, according to Ignatieff (1998, p.119-20) “seek...to ensure that warriors conform to certain basic principles of humanity, the chief principle being to spare civilians and medical personnel.”

19 A similar international example lies in the practice of granting “diplomatic immunity” to embassy staff and their families. According to Shulgan (1999, p.A4), “diplomatic immunity is a reciprocal practice that allows Canada’s diplomats to safely represent the country throughout the world.” The immunity of diplomats, like the immunity of non-combatants during war, is subject to a written agreement. The 1961 Vienna Convention on Diplomatic Relations says that “[t]he person of a diplomatic agent shall be inviolable,” and “[h]e shall not be liable to any form of arrest or detention.”
theory that takes seriously what so many critics have tossed aside with disdain. Its goal is to ask, “What would a well-worked out theory of moral conventions look like?”

Two further points about the method employed here warrant mention. First, this Thesis makes use of methods drawn from the branch of rational choice theory known as game theory. Game theory is the study of strategic interaction; that is, it is the study of situations in which the outcome of action depends upon the combination of various agents’ actions, and in which rational decision making requires an assessment of how one’s own potential actions might interact with other people’s actions. The strategic decision making with which game theory concerns itself is in contrast with decision theory’s focus on parametric decision making, “in which the actor takes his behaviour to be the sole variable in a fixed environment” (Gauthier 1986, p.21). As Gauthier (1986, p.4) notes, game theory analyzes “outcomes in relation to sets of choices, one for each of the persons involved in bringing about the outcome.” Game theory defines any situation in terms of agents, the options available to them, and possible outcomes.

Tools derived from game theory have found their way into an incredibly broad range of disciplines, several of which are drawn upon in this Thesis. Fields that have drawn upon game theory include at least economics (Kreps 1990), evolutionary biology (Maynard Smith 1982), moral philosophy (Gauthier 1986), political science (Hardin 1982), sociology (Kollock 1998), business ethics (Aoki 1984), and psychology (Colman and Wilson 1997).

One of the great benefits of game theory lies in its ability to produce simple schematic models of complex human interaction. Such models allow us to formulate micro-theories of particular situations. That is, they allow us to give abstract characterizations that help us make sense of the way people interact in specific types of circumstances. Game
theoretic models also allow us to see abstract similarities between apparently disparate situations or problems. It is by means of a game-theoretic model, for example, that Danielson (1992, p.6-8) is able to find a common strategic structure in three issues as apparently different as the decision whether to drive or take the bus to work, decisions about national energy policy, and the behaviour of computer hackers.

Like any models, the models used by game theory are simplifications intended to elucidate complicated ideas. Game theory begins with simplified characterizations of agents' preferences (usually), and from there determines the likely outcome of agents' interaction. As Elster (1989b, p.28) puts it, game theory is "a framework for understanding human interaction," and is "more akin to logic than to an empirical science." Like logic, game theory produces no substantive conclusions on its own. It is a tool of reasoning, and the conclusions we reach by using it are only as good as the assumptions we begin with.

The final methodological point to note is the theory of value assumed throughout this Thesis. Although it is the theory of value assumed by our most well-developed social science – economics – it is at odds with most modern moral theories. Although a full theory of value is beyond the scope of this work, the basics – borrowed from David Gauthier's account of value – can be stated briefly. The value theory behind this Thesis is one which holds that all value is dependent upon the affective relationships of the individual, and is therefore both subjective (as opposed to objective) and relative (rather than absolute).

First, that value is subjective means that the ascription of value is taken to be attitudinal, not observational (Gauthier 1986, p. 25). That is, a thing being valuable depends upon some sentient being have a positive attitude toward it. Gauthier (1986, p.46) puts the point quite simply: "Value...is a measure of preference." Thomas Hobbes is perhaps the
classic exponent of the subjective view of value. Hobbes (1651, p.24) writes, “For these words of Good, Evill...are ever used with relation to the person that useth them: There being nothing simply and absolutely so; nor any common Rule of Good and Evill, to be taken from the nature of the objects themselves; but from the Person of the man.” For the value subjectivist then, nothing is ever valuable *simpliciter*; it is always valuable *to someone*. According to Gauthier (1986, p.56), we should decide “whether any adequate account of our experience or environment must refer to objective value” by considering “whether any reference to objective value occurs *in the best explanation* we can provide for our actions and choices” (emphasis original). The value *objectivist*, according to (Gauthier, p.55) “…holds that any adequate account of our experience or our environment must refer to value or to the valuable as being independent of sentient beings and their affections.” Gauthier (p.56) suggests instead that “…the only serious candidate for an explanatory schema for human action is...[one according to which] choice maximizes preference fulfilment given belief. Objective value plays no role in this account.” Gauthier continues: “Objective value, like phlogiston, is an unnecessary part of our explanatory apparatus, and as such is to be shaved from the face of the universe by Ockham’s razor.”

Not only is value to be taken here as subjective, but also as *relative*. On the relative conception of value, according to Gauthier (1986, p.50), “…each person has his own good (and bad), and...the goods of different persons are not parts of a single, overall good.” Note that to hold that value is individualistic in this sense does not imply either ethical or psychological egoism; that is, it does not imply that individuals care only about their own well-being. As Gauthier (p.51) puts it, “A state of affairs that is good for a single individual is then good from the standpoint of each person who prefers or would prefer the
enhancement of that individual’s well-being.” In opposition to the relativist conception of value is the absolutist conception which holds, according to Gauthier (p.50) that “…values are the same for all persons, or for all sentient beings.”

As noted above, most modern moral theories are at odds with this subjective and relative theory of value. Utilitarianism, for example, makes the claim that the happiness of any given individual is a good thing simpliciter; that is, it is to be regarded as a good by every rational being. Thus for Utilitarians, value is subjective but absolute. Kantians hold that the respect due to all rational agents finds its source in their inherent value. Their theory of value must then be both objective (since it makes no mention of anyone’s preferences or affective states) and absolute (since it holds rational agents to have value for everyone). For a fuller defence of the individualist theory that takes value instead to be both subjective and relative, I refer the reader to Gauthier (1986)’s Chapter II section 4.

The Shape of the Project:

This project has three main components. I begin by presenting a positive theory of conventions, arguing that conventions do in fact structure our social interactions and individual choice in important ways. To dismiss conventions as mere historical accidents is to underestimate the crucial moral work they perform. This positive theory is presented in Chapters 2, 3, and 4. I then outline – in Chapters 5, 6, and 7 – a normative theory of conventions, arguing that existing conventions impose real obligations and that new conventions create new obligations. Finally, in Chapters 8 and 9 I present a practical theory of conventions. This part of the project will involve looking at conventions as a moral
suggesting that we can build new conventions to remedy existing social problems, and goes some way toward developing a methodology for doing so.

I begin with a brief review in Chapter 2 of the ways in which conventions and other 'soft institutions' are dealt with in recent work in moral and political theory. This literature is surprisingly sparse, and the terminology it uses varies enormously. Given the variety of such soft institutions (conventions, norms, traditions, etc.), some care must be taken to ensure that conventions are indeed the most fertile ground for attention by moral theorists. Strict ethical relativists, for example, hold instead that cultural norms are the appropriate social structures on which to focus our attention.

Next, in Chapter 3, I examine the ways in which conventions structure the choices with which agents are faced. I note that conventions have both epistemic and motivational implications. On the epistemic side, I argue that particular conventions serve as shared, general mental models of those agents with whom we interact. Next, I examine two different motivational models to explain how it is that conventions affect which actions we choose. I point out that actions taken by agents following conventions seem to be, in the parlance of game theory, actions 'out of equilibrium.' I argue that a departure from traditional game theory is in order, and argue in favour of the multi-dimensional theory of rational choice proposed by Joseph Heath.

Chapter 4 examines the moral significance of conventions through the lens of the literature on public goods and collective action. Here, I argue that moral constraints in general constitute public goods, and hence represent solutions to collective action problems, and are hence conventional in nature.
In Chapter 5 I present the core arguments for a normative theory of conventions. The chapter begins with David Gauthier’s contention that mainstream normative theories face a motivational crisis. I accept Gauthier’s general argument to the effect that it is rational – instrumentally useful – for agents to dispose themselves to at least some forms of constraint. I reject, however, Gauthier’s solution to the compliance problem, along with his argument in favour of specifically impartial constraint. I propose instead that conventions constitute a reasonable set of constraints for boundedly-rational agents like us. In this chapter, I also deal with worries regarding the apparent conservatism of a theory that finds normative force in existing patterns of behaviour.

Chapter 6 tackles the difficult issue of asymmetrical conventions. I point out that some conventions will be asymmetrical in that they reflect, rather than mitigate, power imbalances between the parties involved. Other asymmetrical conventions – i.e., ones better for some parties than for others – will none the less be stable either due to the fact that all available egalitarian outcomes are Pareto-inferior or because they constitute game-theoretic equilibria. If conventions have moral force, then my convention-based theory provides justification for situations that look, from the point of view of everyday morality, unjust. I argue that the egalitarian bent in everyday morality and in moral theory is a bias, and that therefore the propensity of the conventionalist approach to integrate rather than reject asymmetrical relationships may be a virtue, rather than a failing.

In Chapter 7, I examine further the degree to which a convention-based normative theory would diverge from the tenets of everyday morality. I concede that this divergence is significant, but argue that a conventionalist normative theory can have a legitimate role to play given suitably modest aspirations. I propose three lines of defence. First, I suggest that
arguments from convention deserve a spot in the short list of considerations utilized by a pluralistic approach to applied ethics, along with utility, rights, and virtues. In this regard, conventionalism may fill the gap left by the apparent failure of social contract theory to show itself to be useful in applied domains. I suggest that conventionalism has many of the virtues, and lacks some of the failings, of traditional social contract theory. Second, I point out the role of moral sentiments in filling at least part of the gap left between commonsense morality and the obligations that a convention-based approach can sustain. Third, I argue that it may be possible to see a convention-based theory as fully credible on its own, given a more realistically modest estimation of what morals, and moral theory, can accomplish.

Chapter 8 moves from theory to practice. Here, I examine a domain – professionalism – that is subject to a range of moral standards that seem clearly conventional in nature. I propose a meta-ethic of professional obligation that differs radically from the usual principle-based and virtue-based models in suggesting that professional standards find their source in conventions among professionals. I argue that a convention-based model provides much-needed grounding to discussions of the limits of professional obligation and professional autonomy. This model also takes seriously, without endorsing, the power imbalance inherent in the professional-client relationship.

Chapter 9 is the Conclusion of this Thesis. Here, we recap briefly the ground covered, and evaluate the extent to which we have met the objectives laid out in this Introduction, and discuss how we can apply what we know about conventions to ameliorate problematic situations.
Chapter 2:

What are Soft Institutions?
In this Chapter, we will discuss the concept of a ‘soft institution,’ and the need for a moral theory which takes seriously the role which soft institutions have in shaping our moral lives.

The core insight that has made social contract theories so attractive to so many moral and political theorists is arguably the recognition that there is something irreducibly strategic about morality. In this context, the word ‘strategic’ signals the way in which various agents’ actions inevitably interact. In social situations, agents do not get to choose outcomes; rather they choose actions (or strategies), which interact with actions taken by other agents to produce outcomes. A simple example will illustrate this fact: if you and I approach each other on a narrow road, I cannot on my own ensure that we do not collide. If I drive on the left and you on the right, we will collide (since your right is my left). But if I drive on my right and you drive on your right, our actions will be coordinated so as to avoid collision. My action (“drive on the right”) combined with your action (“drive on the right”) produces a particular outcome (i.e., we pass each other unscathed). So the strategy I should choose depends on what I expect you to do, and vice versa. Social contract theory recognizes that the limits I place on my own behaviour should vary according to the limits you are willing to place on your behaviour.

Moral rules, according to social contract theory, are roughly those rules that are needed to bring order to social interaction – or rather, those rules that make recognisably social interaction possible at all. Yet the apparent charms of social contract theory are often offset by worries about the theory’s grounding in dubious counterfactual claims about what ‘rational’ agents would choose in ‘ideal’ situations. Indeed, social contract theory runs so distant a third place among moral theories that most writers in modern fields of application
such as bioethics, business ethics, and environmental ethics accord it no role at all. Writers in those fields typically rely instead on some variant of utilitarianism, some form of rights-based theory, or some combination of the two. Yet modern utilitarian and rights-based theories, though growing ever more complex and subtle, still evidence a striking lack of concern for real social forces and motivations.

What is desperately needed, it seems, is a moral theory that takes account of the inherently strategic aspect of moral obligations, but which avoids social contract theory’s doubly worrisome appeal to hypothetical consent among equals. Although there currently is no such theory, a handful of theorists in several fields have made contributions that promise to foster its development. I refer here to the work done on the nature of social order by philosophers such as Allan Gibbard and George Mavrodes, political scientists such as Russell Hardin, Robert Axelrod and Jon Elster, and economists such as Robert Sugden, on the norms and conventions – what I call “soft institutions” – that shape social interaction. We can also plausibly include fads, customs, traditions, cultural biases, practices, and so on, among soft institutions, but for our purposes here the two most important sorts of soft institutions are norms and conventions.

That norms and conventions should be thought of as “institutions” will strike some as obvious, and others as counter-intuitive. Clearly they have the features we see in other

---

1 See, for example, the most widely respected bioethics text, (Beauchamp and Childress 1994).
2 See, for example, (Velasquez 1992).
3 See, for example, the introduction to (Pierce and VanDeVeer 1995). See also the dismissal of contractarianism by (Regan 1985), reprinted in that same volume.
4 I return to this topic in Chapter 7.
institutions: they motivate people, they structure interaction, they attract allegiance, etc. Yet the fact that calling them such will strike some as odd implies that the name is not redundant. I call them ‘soft’ for the simple reasons (roughly) that they tend to be rather more amorphous than most institutions (the conventional speed limit on a given road is less certain than the legal limit), and their injunctions less reliably enforced (since it can be costly to enforce norms if that’s not your job). Petit (1990, p.725) agrees with this usage of the term ‘institution,’ stating that norms are “an important species of social institution, on a par with conventions, customs, laws, and other brands of established regularity” (emphasis added). Such soft institutions can be compared to full-fledged legal, governmental, religious, and corporate institutions. Elster (1989, p.147), on the other hand, defines an institution as a rule-enforcing mechanism that “governs the behavior of a well-defined group of persons, by means of external, formal sanctions.” He writes that this is in contrast to norms, “which enforce rules by external, informal sanctions....” We will not quibble with Elster on this matter of terminology. For our purposes, we will apply the term “institution” to a somewhat broader range of social forces. We will do so for three reasons, each having to do with signaling particular aspects of norms and conventions that are particularly important to the project at hand. First, this terminology is justified as a way of pointing to the reality of norms and conventions. Norms and conventions are real social phenomena that exist as matters of objective fact, independent of the opinions of particular persons. Second, norms and conventions are institutions in that they are in the broadest

5 (Mackie 1977, p.80) treats as institutions anything that is “constituted by many people behaving in fairly regular ways, with relations between them which transmit and encourage and perhaps enforce those ways of behaving.”

6 I return to the problem of enforcing conventions in Chapter 4 below.
sense artefacts; that is, they are in some sense *instituted* by human action. Finally, we apply the label “(soft) *institution*” to norms and conventions in order to highlight their more or less *structured* influence on behaviour.\(^7\) That is, they constitute ‘law-like’ regularities of behaviour.\(^8\)

The difference between a soft institution and a more formal institution can be illustrated by comparing the informal social enforcement of etiquette with the formal enforcement of the criminal law. The social enforcement of etiquette imposes a significant degree of order on our activities in many domains, though its demands are nowhere codified.\(^9\) Etiquette is an institution in the sense that it is a system that orders the behaviour of a range of individuals, and in that its demands, and the penalties for failure to meet them, are generally well-defined and well known. Etiquette, for example, precludes (in most circles these days) smoking in someone else’s home without first asking permission. The individual who breaks this unwritten rule is likely to garner looks of disapproval or exaggerated coughs, and may not be invited back. Etiquette also has enough structure as a system that most people can differentiate the demands of etiquette from the demands of, for example, laws, moral injunctions, and corporate policies. We know not to clip our fingernails at friends’ dinner tables, despite the fact that there is no law against doing so,

\(^{7}\) In a recent debate in the pages of the *Cornell Law Review*, Macey (1997) and Rutten (1997, p.1164) agreed that “norms are part of the web of *institutions* – simply one more way that people order their social world.” (emphasis added).

\(^{8}\) (Rawls 1955) uses the term “practice” in a similar way. In the 2\(^{nd}\) footnote to that paper, he writes, “I use the word ‘practice’ throughout as a sort of technical term meaning any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure. As examples one may think of games and rituals, trials and parliaments.”

\(^{9}\) The demands of etiquette can, of course, be written down, but written accounts of those demands will seldom be accorded significant authority.
that the venues in question are outside the purview of any corporate policies, and that no one would think us immoral for doing so. But etiquette (like other customs as well as traditions, fads, etc.) is a soft institution in the sense that its directives (unlike legal standards, church doctrines, and corporate policies) are typically not codified and not subject to authoritative description.

As indicated above, we can count among soft institutions fads, traditions, practices, and so on. For our purposes here, the two most important sorts of soft institutions are norms and conventions. The theorists writing in this area vary to a significant degree in their use of the terms ‘convention’ and ‘norm,’ and indeed in their estimation of the relative importance of the two concepts. In this Chapter, I survey this varied literature. The authors discussed here differ greatly in the ways in which they define the soft institutions concerned, and in the theoretical use to which they put soft institutions. It is in particular worth noting that the literature referred to here is almost exclusively descriptive in nature. The literature on norms and conventions has focused on such questions as why (what I have called) soft institutions exist, what causes a given soft institution to emerge, and what allows it to persist. This literature has generally ignored questions of the prescriptive force that should be accorded to soft institutions. Yet it is precisely the prescriptive force of soft institutions that needs to be explored if we are to supplement existing normative theories. My discussion here will be motivated in part by the question of which soft institutions we should focus on if we wish to examine their prescriptive force, and their usefulness as a moral category.

This Chapter has 4 sections. In Section 1 I begin by discussing the literature on norms. In Section 2, I move on to a discussion of the (relatively few) authors who have
focused on the notion of a *convention* as it applies to ethics. In Section 3 I touch briefly on some related categories of soft institutions that enrich the set of social structures to which we can appeal in building our socially-savvy moral theory. In Section 4, I argue for the greater moral relevance of conventions over other categories of soft institution.

**Section 1: What is a Norm?**

The literature on norms, in comparison to that on conventions, is vast and varied. There is, for example, an extensive body of literature on the interaction of law and norms.\(^{10}\) The language of norms is also endemic to the work of sociologists.\(^{11}\) For the purposes of this Thesis, I shall ignore these and focus primarily on those writers whose work is of most relevance to moral theory, as I see things.

Let us begin by offering a thumbnail definition of "norms" that is consistent with the way most moral and political theorists use the term. Roughly, a norm has the characteristics that it is both a) a social (i.e., group-level) regularity of behaviour, and b) a motivating force for the individual (i.e., people generally think that they – and others – ought to follow it).\(^{12}\) Both of these aspects are at least implicit in the work of all of the authors discussed here, although there are interesting differences in the relative weight placed on those two characteristics. Familiar norms include limitations on the use of vulgar language in public, the avoidance of public nudity (in North America), and the concept of *noblesse oblige*. Some (mostly) outdated norms include the idea that a woman’s place is in

---

\(^{10}\) For a comprehensive review of this literature, see (McAdams 1997).

\(^{11}\) For examples, see the first footnote of (McAdams 1997).

\(^{12}\) I acknowledge that this usage is at odds with the alternative usage of the term, according to which a “norm” is purely prescriptive and can exist independent of whether or not it constitutes a regularity of behaviour.
the home, the idea that children should be seen and not heard, the idea that people with dark
skins warranted less respect than people with lighter skins, and the idea that a gentleman
ought to offer his seat on a bus to a lady. Also included are Sicilian standards of vengeance,
the *lex talionis*,\(^\text{13}\) and the belief in universal human rights. In contrast, the principle of
utility is not a norm, in that it does not in fact constitute a group-level regularity of
behaviour: some aspire to it, but it is not an accurate description of our collective conduct.
Nor is gossiping a norm: it may be accurately described common behaviour, but it is not a
behaviour in which people think they *ought* to participate.

As a further characteristic of norms, note that norms are usually taken to be matters
of social contingency, rather than for example biological necessity. Thus the desire for food
or for sexual gratification is not a norm in the sense discussed here.\(^\text{14}\)

Let us look first at the regularity aspect of norms. We begin our discussion of norms
with political scientist Robert Axelrod. Axelrod (1997) is explicit in treating norms as an
observable group-level characteristic. He notes (p.45) that we often observe “large numbers
of individuals and even nations” displaying a high degree of regularity of behaviour.
Sometimes, that regularity will be explained in terms of the existence of some central
authority acting as an enforcement mechanism. In other cases, however, there is no such
central authority (or at least no central authority policing *this* behaviour). In such cases,
writes Axelrod (p.45), “we tend to attribute the coordinated behavior and the resulting
regulation of conflict to the existence of norms.”

\(^{13}\) i.e., the “eye for an eye” standard of retributive justice.

\(^{14}\) According to (Snare 1991, p.216) it is a crucial part of Hume’s account of behaviours governed
by convention that there be “no ‘natural’ motives sufficient to explain...compliance.”
Axelrod offers (p.47) the following behavioural definition of a norm: “A norm exists in a given social setting to the extent that individuals usually act in a certain way and are often punished when seen not to be acting in this way.” Since Axelrod’s interest is in computer simulations of the spread of norms, it is not surprising that he focuses on what is observable (i.e., regularity) and on simple causal properties (e.g., the imposition of sanctions). Indeed, his definition embodies two sorts of regularity: regularity in following the norm, and regularity in enforcing the norm. Note that Axelrod’s definition – as presented – makes no causal claim: it does not imply that the regularity is a result of the punishment meted out to those who deviate. But in his discussion Axelrod does, of course, assume such a causal relationship: indeed, he produces a computer simulation illustrating that punishment can be an effective way of stabilizing cooperation. Note also, however, that Axelrod’s insistence that norms be enforced is too restrictive. Axelrod uses enforcement as a behavioural indicator of the prescriptive force that is part of the concept of a norm. Although his interest in observable characteristics requires him to make this move, it results in too narrow a definition. If we insist on observing acts of punishment before we call a pattern a “norm,” we will end up excluding what look like central cases of norms. For instance, one of Axelrod’s own examples of a norm, namely (p.46) the use of gender-laden language, seems not to fit with his definition: presumably no one was punished in past days for using the pronoun “she” to describe a generic agent. The prescriptive force of a norm such as this, it seems, is permissive rather than obligatory. That is, the fact that gender-laden language was the norm is indicated by the facts that a) it constituted a regularity of behaviour and b) nobody thought this regularity inappropriate.
The requirement that we actually observe acts of punishment is overly restrictive in another way: some behaviours may constitute social regularities that individuals see as prescriptive, without there actually having been instances of punishment for deviation. Axelrod focuses on punishment, presumably because so many central cases do involve social sanctions of some sort, and because punishment is relatively easy to model. But threats often work just as well as actual punishment. For instance, it may be that a given norm is so well entrenched that everybody knows that they would be punished if they were to deviate, even though nobody had ever in fact deviated. It might be the case that the proscribed behaviour is so outrageous that nobody had ever dared test the waters by deviating. What is perhaps needed, then, is a counterfactual definition: a norm exists in a given social setting to the extent that individuals usually act in a certain way and are likely to be punished when not seen to be acting in this way. Yet even this is too restrictive, for some forms of behaviour constitute regularities but are taken as prescriptive for moral reasons. For example, our society presumably has a norm forbidding the torture of babies. Indeed, the torture of babies is all but unknown in our society. However, this regularity of behaviour is presumably best understood as resulting from a shared feeling that torturing babies is wrong. And this could be true even in a society that did not believe, for example, that punishing wrongs is ever justified. Thus norms are regularities of behaviour that individuals understand as prescriptive, for whatever reasons.

This prescriptive nature of norms is highlighted by philosopher Allan Gibbard (1990a, p.791) who instructs us to think of a norm “as an imperative saying what to do in some kind of situation, or how to feel.” Gibbard treats the term “norm” as roughly equivalent to the term “rule.” That is, for him a norm is roughly any standard to which an
individual holds herself. For him, that a norm is shared is a contingent matter: it need not be shared in order to *be* a norm. Gibbard’s main concern is to ask why it is that norms so often motivate us. What is important to Gibbard (p.791) is that norms influence individual behaviour: “...a person who accepts a norm tends to abide by it when he encounters a situation to which it applies.” Although his focus is on the motivational aspect of norms, the point of this focus for Gibbard is to offer an explanation of how it is that norms manage to motivate us in ways that facilitate the kinds of cooperation upon which efficient outcomes depend.

The motivational aspect of norms is also the focus of political scientist Jon Elster. For Elster, a norm as such need not be widely shared: norms may explain regularities of behaviour, but they do not *constitute* them. For shared norms, Elster reserves the term “social norms.” For norms to be classified as social, on Elster’s (1989a, p.99) account, they must be “(a) shared by other people and (b) partly sustained by their approval and disapproval.” Thus Elster’s term “social norm” coincides with what we have called “norms” more generally. One of the things that distinguishes Elster is that on his (1989a, 1989b) account, norms are action guides that are essentially deontological in nature. That is, norms according to Elster (1989b p.113; 1989a, p.98) are never outcome-oriented in the sense that they never justify action on the basis of individual or group benefit, even though they may (1989b, p.117) in fact *produce* such benefits. They are simple rules, such as ‘Insults to one’s honour must be avenged,’ which direct action without reference to some further goal.  

---

15 When a norm is explicitly goal-oriented, Elster (1989a, p.132) calls it a “moral norm.”
Elster's primary interest, however, is in the relationship between norms and interest. He points out (1989a, p.147) that while some norms are beneficial to individuals, and others are beneficial to groups, some norms are bad for all concerned. Thus, (1989a, p.125) there is "no single end – genetic, individual or collective – that all norms serve and that explains why there are norms." Elster argues (1989a, p.151) that while some cases of adherence to norms might be explainable in terms of self-interest, it is more plausible to think of both norms and self-interest as influencing behaviour more or less independently.

Political scientist Russell Hardin (1995, p.108) agrees with Elster’s claim (or rather stipulation) that norms are never outcome-oriented. Unlike Elster, however, Hardin is concerned to explain norms – particularly norms of ethnic and nationalistic identification – in terms of the hidden rationality of adhering to them. That is, Hardin claims that adherence to norms can be understood as rational if the interplay between individual and group benefit in particular cases is properly understood.

Philosopher Philip Pettit (1990) also examines norms primarily from the perspective of motivation. For him (1990, p.725), "norms reinforce certain patterns of behavior...by representing those patterns as peculiarly desirable or obligatory." For Pettit (and us), conventions are merely a species of norms. Indeed, his own discussion of the preconditions for the emergence of certain kinds of norms draws heavily on David Lewis’s near-canonical discussion of the preconditions for conventions (to which we return below). Pettit’s definition of a norm is very much like Lewis’s definition of a convention. Pettit (p.751) writes:

"A regularity, R, in the behavior of members of a population, P, when they are agents in a recurring situation, S, is a norm if and only if it is true that, and it is a matter of common belief that, in any instance of S among members of P,
1. nearly everyone conforms to \( R \);
2. nearly everyone approves of nearly anyone else's conforming and disapproves of nearly anyone else deviating; and
3. the fact that nearly everyone approves and disapproves on this pattern helps to ensure that nearly everyone conforms.” (emphasis original)

Philosopher Edna Ullmann-Margalit (1990) focuses primarily on the regularity aspect of norms, and particularly on how norms change over time. Her (p.756) “working definition” of a norm is “a regularity such that people generally conform with it, and, moreover, they generally approve of conformity to it and disapprove of deviance from it.” She describes a category of norms, “coordination norms,” that very closely resemble conventions as described by other authors.

Philosopher Howard Margolis (1990) tackles directly the ambiguity of the word “norm.” Margolis (1990, p.821) points out that “[i]n one sense a norm is just whatever usually occurs. In another sense (or cluster of senses) a norm signifies a standard of proper behavior.” He further distinguishes between four categories of norms: aspirational norms (judgments of ideal behaviour), pragmatic norms (standards which individuals actually expect each other to follow), statistical norms (actual behaviour), and personal norms (by which an individual judges her own behaviour). Margolis’s “personal norms” are most like norms as dealt with by Elster: they direct action, and need not be shared across individuals. To the extent to which a standard is a personal one, not shared or preached, it will not be of interest to us here. The “statistical norms” dealt with by Margolis are important to mention by way of contrast. They point to one usage of the term “norm” – as seen for example in such exasperated exclamations as, “Violence among teenagers seems to be becoming the norm!” – that is very different from the “normative” sense in which we use the term here. “Aspirational norms” too will be largely ignored here, except to the extent to which they
are shared and affect behaviour. When a group of people aspire to the same norm (e.g., “Do unto others…” ) it may have an effect on their collective behaviour, even if it does not accurately describe their behaviour. The norms of interest to us here, then, are Margolis’s “pragmatic norms” – i.e., standards which do affect behaviour and which individuals actually expect each other to follow.

Section 2: What are Conventions?

The exact nature of conventions is a matter of some dispute. That is, the use of the term “convention” varies across theorists. Many authors (particularly those who focus on norms) refer to conventions in a way that minimises their importance relative to norms. Axelrod (1997, p.62), for example, writes that a convention “has no direct payoffs one way or the other (such as wearing a tie for men)…." and he contrasts conventions with cooperative acts, “the violation of which leads to injury to others (e.g., queuing for service).” Among authors who take conventions more seriously, two general characteristics unite various authors’ conception of a convention. First, conventions form a subset of the set of norms. That is, all of the authors discussed here regard conventions, too, as group-level regularities of behaviour that individuals take to be in some way prescriptive. Second, conventions involve coordination for mutual benefit. That is, though they disagree as to whether conventions are purely coordinative, or merely involve a partial desire to coordinate, the authors here agree that conventions must embody some shared interest. Note that this differentiates conventions from the larger set of norms, which as noted above need not benefit – nor be seen as benefiting – anyone.
We could do worse than to begin our survey of the literature on conventions with the philosopher whose name is most often associated with a conventionalist approach to ethics, namely David Hume. Hume’s conventionalist analysis covered such important moral concepts as property, justice, and promising. Hume’s conception of a convention, as he uses it in his *Treatise*, is of a regularity of behaviour, sustained by each agent’s having both an interest in coordination and an expectation that other agents will conform. For Hume, conventions are artifices by which humans overcome the disadvantages of their circumstances. These unfortunate – but remediable – circumstances are constituted by a combination of limited (but not non-existent) natural sympathies on one hand and the scarcity (and uncertainty) of certain desirable goods on the other. According to Snare (1991, p.217) a convention exists, for Hume, when a given regularity of behaviour is best explained by people having the relevant motivations and expectations (despite the temptations to unilateral deviation). On Hume’s account, conventions occur when individuals in such situations regulate their behaviour because they understand that doing so is in their mutual interest. Hume (*Treatise* III.II.V, p.490) writes, “When this common sense of interest is mutually express’d, and is known to both, it produces a suitable resolution and behaviour.”

Note that on Hume’s conception, a convention solves a particular kind of coordination problem, namely a *mixed-motive* problem. For Hume, conventions do not solve problems of pure coordination, such as the problem of choosing on which side of the road to drive. Thus his famous example of the two rowers is an unfortunate anomaly, since

---

For example: (*Treatise*, III.II.V, p. 516) “…a promise wou’d not be intelligible, before human conventions had establish’d it.”
it involves what is clearly a pure coordination problem.\textsuperscript{17} Although Hume emphasizes the cooperative aspect of convention, his explanation of adherence to conventions makes it clear that the situations with which he’s concerned do not involve an \textit{unmixed} interest in coordination. Pure coordination conventions solve a particular kind of problem, namely an epistemic one. When it comes to choosing lanes on the highway, our interests do not diverge: simply \textit{knowing} what you intend to do is enough to allow me to choose among actions among which I am otherwise indifferent. There can be no conflict here, only confusion. As noted above, Hume’s discussion of convention begins with conflict: we are naturally greedy, and have no reason to trust each other. Limitations on trust and security put serious constraints on human well-being. Conventions help, but they do not utterly eliminate the causes of the problem. For example, he notes (\textit{Treatise}, III.II.V, p.521) that given a convention of promise keeping, a promisor will keep his promise even after he reaps the benefits of having made the promise, but he will do so because he sees “the consequences of his refusal.”

Next, let us look at what is generally taken to be the seminal contemporary work in the area, namely philosopher David Lewis’s 1969 book, \textit{Convention}. Lewis’s (p.208) informal definition of a convention is that conventions are “regularities of behavior, sustained by an interest in coordination and an expectation that others will do their part.”\textsuperscript{18} Lewis (p.3) claims that this account of convention is Humean. But Lewis’s (p.45) definition of a convention is a somewhat restrictive one: on his account, in order for a regularity to be

\textsuperscript{17} Harman (1983) takes Hume’s rowers to be involved in a mixed-motive game, on the assumption that each of them has a preferred \textit{rate} of rowing: each would prefer to coordinate on \textit{her} preferred rate, but each values coordination (or compromise) even more.

\textsuperscript{18} See (Lewis 1969, p.42) for a more formal definition.
a convention, it must constitute a coordination equilibrium. That is, on Lewis’s account, for a regularity to constitute a convention, it must be true not only that no one is motivated to deviate unilaterally, but also that no one wants anyone else to deviate. Any conflict of interests, according to Lewis (1969, p.46), prevents a regularity from being a means of reaching a coordination equilibrium, and thus precludes the possibility of true convention. Thus the element of conflict which was so clearly a part of Hume’s use of the term is ruled out under Lewis’s usage. Lewis (1969, p.46) defends this usage simply on the grounds that “the presence of substantial conflict makes a disanalogy between [such cases] and other clear cases of convention, and makes an important analogy between [such cases] and clear cases of nonconvention....” Of course, in order to make this claim, he needs to deal with the conflict that is so apparent in the range of problems that Hume says are solved by convention. For instance, one of the examples that Lewis borrows from Hume is the division of land. But in order to make the division of land into a game of pure coordination, Lewis asks us (p.48) to assume “that nobody ever thinks it worth the trouble to try taking over the use of some land from another....” As argued above, Hume’s account of convention involves, for the most part, situations in which our interests do conflict. Thus – pace Lewis – it is only in a rough sense that Lewis’s account of convention can be called “Humean.”

In Lewis’s schema, conventions constitute a subset of the class of norms. For him as for us, norms (p.97) are “regularities to which we believe one ought to conform.” He argues that whenever a regularity constitutes a convention, it will – given the convergence of

---

19 Skyrms (1996, p.69) uses more precise game-theoretic language to state that Lewis’s account “takes a convention to be a robust Nash equilibrium of a coordination game.”
interests inherent in coordination equilibria – be the case that a) my conforming would be to my own advantage, and b) my conforming will be to others’ advantage. And since, he notes (p.98), we generally presume that we ought to do things that answer both to our own and to others’ preferences, we therefore have reasons to do any action that conforms to a convention. Conventions, therefore, are a species of norms.

For economist Robert Sugden (1990, p.779), conventions are “stable patterns of expectations” such that it is “rational for each person to hold these expectations and to act in a way that sustains these expectations in others.” For Sugden, the most significant aspect of conventions is that a given convention is always one of several possible solutions to a game. The main puzzle, from Sugden’s point of view, is how it is that one particular way of behaving can come to be socially pervasive, without conscious human design. He argues that the existence of a convention can provide a point of salience – a bit of shared experience – around which to coordinate action in those situations in which rational choice is (otherwise) indeterminate.20 (We will return to this question of the indeterminacy of rational choice in Chapter 4.)

Sugden (1989) associates the idea of a convention with the concept, borrowed from evolutionary game theory, of an evolutionarily stable strategy (or ESS). An ESS (1989, p.91) is a “pattern of behavior such that, if it is generally followed in the population, any small number of people who deviate from it will do less well than the others.” Sugden (1989, p.91) defines a convention as “any ESS in a game that has two or more ESS’s.” Sugden notes (1989, p.93) that once enough people follow a particular convention (for

---

20 Sugden attributes this idea of “prominence” or “salience” to Schelling. In particular, he cites (Schelling 1960, p.54-8).
whatever reason), then "a self-reinforcing process is in motion." While Sugden, like Lewis, focuses on conventions as equilibria, Sugden notes (1989, p.96) that he differs from Lewis in moving beyond games of pure coordination. On Sugden's account conventions are equilibria, but they need not be coordination equilibria. That is, given the knowledge that others are likely to act according to convention, any given agent will do best to act according to convention as well. But (and this is where Sugden differs from Lewis) it need not be the case that the agent actually wants others to adhere as well. Thus Sugden makes the important move to conventions that mediate mixed-motive games.

Political scientist David Brown, whose stimulating book, *When Strangers Cooperate* (1995), led me to this topic, sees conventions in an enormous range of social activities. As Brown points out (p.62), there are many situations in which "[w]ithout deliberation or instruction, people somehow manage to cooperate." He argues that such cooperation is possible because conventions structure social interaction. Brown (p.17) uses Lewis's simplified definition of a convention, defining them simply as "...regularities of behavior, sustained by an interest in coordination and an expectation that others will do their part." Unlike Lewis, however, Brown's interest (1995, p.20) goes beyond games of pure coordination to situations in which "players prefer the same outcome but have other attractive alternatives so it is by no means certain that enough people will coordination their behaviour to produce the desired outcome."  

---

21 Indeed, some of the regularities Brown calls "conventions" (e.g., (p.16) "joining a self-help group") sound more like clichés, and are hard to square with the definition he gives.

22 Some of the examples of conventions that Brown names but does not discuss, such as "flip-a-coin," "hitchhiking," "smoke signals" seem like matters of pure coordination.
Brown’s main focus is on the predictability aspect of conventions. Conventions, on Brown’s account, allow us in certain situations to trust strangers. As Brown points out (p.19), “...one of the primary functions of a convention is to reduce uncertainty” about how others will act. The existence of a convention ensures the individual that others will act in a particular way, and thus that her contribution to the collective effort will not be wasted.

Philosopher George Mavrodes (1990), who explicitly defends a morality based in part on convention, makes use of the concept of a convention without offering much in the way of specification. His account treats a convention as a modus vivendi, a simple mutually beneficial way of getting along. But his use of the term “convention” does suggest a few characteristics that he thinks conventions have. First, conventions can be actively created; they need not represent purely tacit agreement. Further, Mavrodes agrees with Hume that they can be used to remedy situations approximating mixed-motive games. Finally, he points to the way in which acting according to convention implies an interest in reciprocity. All the theorists discussed here, of course, understand conventions in terms of conditional behaviour; but Mavrodes more than the others makes clear that where behaviours are conventional, we are only motivated to consider ourselves to have obligations in that regard to the extent that others act accordingly. For Mavrodes, even the obligation not to kill the innocent is not absolute: it is a purely reciprocal duty.

Finally, philosopher Gilbert Harman (1983) has gone farther than any theorist discussed here to defend the idea that morality is based on mutually beneficial conventions. Harman argues that conventions constitute an implicit bargain between the people involved. He suggests (p.123) that conventions are “normally arrived at tacitly, by mutual adjustment of different people’s behavior, without conscious awareness.” Harman’s main interest is in
the motivational aspect of morality. He suggests that we need to ask what reasons a person
would have to follow a particular moral rule. He submits (p.123) that “the main reason why
a person accepts the principles he or she accepts is that it is in his or her interest to do so if
others do too.” That is, “we accept the moral principles we accept as useful conventions.”

Section 3: Other Relevant Categories of Soft Institutions

The ontology of soft institutions at our disposal can be enriched by a few more
concepts related to norms and conventions.

Axelrod (1997, p.52-55), Elster (1989b, p.120 ff.), and Pettit (1990, p.739 ff.) each
discuss the possibility of higher-level norms mandating the punishment of those who fail to
adhere to a first-order norm. Axelrod (1997, p.52) refers to such a norm as a ‘metanorm.’
Metanorms are posited as a stabilizing mechanism: a given norm is more likely to be
adhered to if those who fail to adhere to it are punished. But (as we will discuss further in
Chapter 4) it is seldom in anyone’s direct interest to punish deviants themselves. Thus
deviants are more likely to be punished if there is a norm – a shared standard – that says
that deviants ought to be punished. How would such a metanorm be sustained? In principle,
it could be sustained through a further norm mandating the punishment of those who fail to
punish those who fail to meet the requirements of first-order norms. This leads of course to
the possibility of meta-meta-norms. Gibbard (1990, p.797) rightly questions the likelihood
of there being an infinite hierarchy of dispositions to punish deviation from norms.

Sugden (1989, p.93-4) also mentions the idea of a families of conventions (what we
might call ‘metaconventions’) by which he means a convention-type (e.g., ‘first come, first
served’) that may serve as a template for an entire class of conventions-tokens (e.g., bank
line-ups, primogeniture,\footnote{Primogeniture is the practice according to which the bulk of a family's wealth is inherited by the first born son.} etc.). Sugden argues that the similarity which a novel situation, “A,” bears to a familiar one, “B,” may lead to the establishment of a convention to govern A bearing a family-resemblance to the convention governing B. Sugden offers this as a partial explanation of how it is that we come to follow one particular convention instead of another, possibly equally good one.

We see then that the literature is rich with descriptions of the soft institutions that shape our moral choices and social interaction. This provides much food for thought, and a respectable set of building blocks out of which to begin constructing the socially-savvy moral theory for which we wished in the introduction to this Chapter.

Section 4: Narrowing Our Interest

As suggested above, my discussion here has been motivated in part by the question of which soft institutions we should focus on if we wish to examine their prescriptive force, and their usefulness as a moral category. It has already been suggested here that conventions are the soft institution on which we ought to concentrate in this regard.

A key difference between norms and conventions, as described by the authors discussed here, is that conventions are sustained in part by a genuine (if limited) interest in coordination. Norms, on the other hand, tend to enforce conformity independent of any intrinsic motivation. This must be counted a disadvantage, both in terms of propagation and stability (since I am less likely to accept and adhere to a rule that does not speak to my interests). If it is true (as I believe it is) that the primary function of morality – that is, the
task we want it to perform – is to overcome some major obstacles to getting along in social situations, then we must be interested in coordination, not just in regularity. This suggests that conventions, rather than norms, are the soft institution on which to focus.

We have defined conventions as regularities of behaviour, sustained by an interest in coordination and an expectation that others will do their part. Conventions, then, as we have noted can be defined in terms of these three key concepts: regularity of behaviour; an interest in coordination; and an expectation of cooperation. The degree to which various conventions will embody these three characteristics will vary, but all three characteristics will always be present to some degree.

Note also that none of these three crucial concepts (regularity, interest, expectation) is adequate on its own to characterize plausible solutions to moral problems. Other social structures or patterns of interaction may have one or more of these characteristics, but not all three. First, customs and fads certainly involve regularities of behaviour, but lack the kind of importance – the interest in coordination – that would give us either a real interest in maintaining them or an ethical expectation that others will do so. Such situations provide little in the way of portable solutions to moral problems, since undermotivated fads are inherently unstable. (Indeed, they are by definition ephemeral).

Secondly, in novel situations (as in situations approximating a true one-shot PD), each of us may have a partial interest in coordination, but no established pattern of behaviour and thus no reason to expect (nor ability to issue threats in order to compel) the other to act in the way we would like. The difficulties in solving such situations are tragically well known.

---

24 Partial, because the PD is a mixed-motive game.
Finally, in some situations we might expect (i.e., demand) a certain kind of behaviour from someone without there being either an established pattern of behaviour or any interest in coordination. In such situations, we might assert that there is a principle that requires us (all of us) to act in a certain way. But such moralizing in situations of pure conflict is unlikely to be persuasive.

Conclusion

I conclude, then, that the relevant soft institution (for moral theorists interested in real social forces) is roughly the Humean mixed-motive convention. The broader category of norms may well serve to explain behaviour in a range of situations. But if our goal is to seek rules that can be taken to have prescriptive force, conventions seem like the more promising soft institution.
Chapter 3:

How Conventions Affect Choices: Knowledge and Motivation
On most conceptions of morality, moral lives are defined primarily in terms of choices. Morality is most plausibly seen as a set of limits on the choices agents make. What role do conventions play in our moral decision making? The most drastic effect which conventions can have on our decision making is to obviate the need for choice. (Few of us think of ourselves as facing a choice when we show up at the bank; we just get into line.) When choice *is* required, conventions can play both an epistemic role and a motivational role. In this chapter, I examine the ways in which conventions affect the choices which agents make. On the epistemic side, I argue that particular conventions serve as shared, general mental models of those with whom we interact. Such being the case, the existence of a convention provides information that must affect strategic interaction. Then I turn to the motivational aspect of conventions. I examine two different ways of thinking about how it is that conventions affect which actions we choose in particular situations. I point out that actions taken by agents following conventions are, in the parlance of game theory, actions ‘out of equilibrium.’ I examine four ways in which traditional game theory might deal with this, and find all four of them lacking. I claim that a departure from traditional game theory is in order, and argue in favour of the multi-dimensional theory of rational choice proposed by Joseph Heath.

**Part 1: Conventions as Shared Mental Models**

Economists Arthur Denzau and Douglass North have written of the importance of shared mental models in understanding social choices and interactions. They argue (1994, p.5) that “[w]e need to develop a framework that will enable us to understand and model
the shared mental constructs that guide choices and shape the evolution of political-economic systems and societies.” In this section, I argue that conventions are an important class of shared mental models, and that as such they simplify important cognitive tasks required for certain kinds of social interaction.

Strategic choices — even in simple, two-player situations — require the agent to have mental models that let her anticipate the actions of those with whom she interacts. Rational choice theory makes the assumption that all agents have accurate mental models of those with whom they interact; that is, it assumes that agents have access to each other’s utility functions.¹ Like any simplifying assumption, this one should be endorsed to the extent to which it makes analysis easier without jeopardizing the predictive value of the model. Friedman (1953, cited in Denzau and North 1994, p.5) notes that the “substantive rationality” model of which this assumption is part need not be accurately descriptive of particular individuals, so long as it is accurately predictive of aggregate behaviour. But the assumption that all agents have accurate mental models of those with whom they interact hides the cognitive costs of such models. Building, testing, and modifying such models is likely to be costly both in terms of time and energy as well as in terms of the costs — which may be extreme — borne by the agent as a result of failures in early, inaccurate versions of the model. This complexity may make individual model-building problematic even in the case of simple two-player interactions.²

As Denzau and North (1994, p. 11) point out, “[a]n even harder choice would involve a...situation with a multiplicity of other agents.” And of course, a large portion of

¹ i.e., they all know the value placed by other agents on particular outcomes.
² Indeed, in one-off interactions, building a model of one’s opponent based on experience is hardly possible.
our social interactions involve more than one other person. If such situations required agents to build separate models of each agent with whom she interacts, the cognitive costs could be prohibitive, or at least would greatly reduce the net gain to be had from interaction. Also, the level of certainty — i.e., the accuracy of our mental models — is likely to drop as the number of agents involved increases. Denzau and North argue that ideologies function as shared mental models, thus simplifying the cognitive demands of social interaction.

Given the costs of building internal models of a multiplicity of other agents, we should expect to see strong selective pressure — either biological or cultural — in favour of any trait that tends to reduce the need for agents to build a separate model for every agent with whom she interacts. As Denzau and North (1994) argue, one way to simplify interaction is for agents to have a mental model that is both generalized across agents and shared among agents. Generalized models simplify the cognitive component of interaction in that they purport to be models of a wide range of agents, eliminating the need for a given agent to maintain a plurality of models. A general mental model is a model not just of one agent but of a class of agents. There must, of course, be a trade-off between generality and accuracy when models purport to apply to individuals who in fact differ from each other.3 Shared models further simplify the cognitive aspect of interaction in that they can be acquired more or less directly from other agents, rather than having to be built from the ground up by each of us. As Denzau and North (1994, p.14) point out, “[t]he world is too complex for a single individual to learn directly how it all works.” Shared mental models,

---

3 This is of course true of all kinds of models. For a discussion of this issue as it pertains to computer modeling, see (MacDonald 1998b).
then, reduce cognitive costs involved in learning about how other agents may be expected
to act.

I suggest that *conventions* can and often do function as mental models in
simplifying the cognitive tasks involved in social interaction. Indeed, in many instances, it
will make sense to model other agents as behaving *conventionally* rather than as acting
*rationally* (where convention and rationality suggest different behaviours). For instance,
when standing in line at the bank, I can anticipate the behaviour of the next person to walk
through the door. I have no idea what cognitive processes will underlie her behaviour, but I
am almost certain what that behaviour will be. I anticipate (based on the model I have
constructed – based on experience – of my fellow Vancouverites)⁴ that the next person to
walk in will act in the conventional way and join the queue. It would almost certainly be
*rational* (i.e., more conducive to her self-interest), however, for her to step straight to the
front of the line instead, given that a) we polite Vancouverites are likely to be caught off-
guard by such audacity, and are likely unprepared to impose more than token sanctions, and
b) her unilateral deviation is very unlikely to destabilize the entire convention. Were I to
model my fellow bank customers as rational agents, rather than as convention-followers, I
would expect each person who enters the bank after me to try to jump the queue. And I
would be wrong every time.

⁴ This is one way for me to anticipate her action, and I think it the most likely one for situations like
this. The alternative is to see *myself* as a model of how she will behave. “If I were her, I’d
probably…” But even if I were to use this second means of anticipating her behaviour, I would still
base my calculations in part on my knowledge of the existence of the relevant convention. For an
overview of the debate between “simulation theory” and “theory theory,” see Gordon and Barker
(1994).
Any choice situation requires an action-outcome mapping. In strategic choice, the actions which other agents may take, and the way they will react, form a crucial element of that mapping. To the extent to which other agents can reliably be expected to adhere to conventions, the existence of a convention will give the agent a strong indicator of how other agents will act. In strategic interaction, the assumption that A will follow a known convention provides B with a clear mapping of her own actions onto outcomes. And knowing that there is a relevant convention allows the individual to anticipate the immediate reaction that non-conformity will engender, and the degree to which her non-conformity imperils future opportunities.

Another way of thinking of the epistemic effect of conventions is to think of the existence of a reliable convention as essentially turning an erstwhile strategic choice into a nearly parametric one. And as Denzau and North (1994, p.9) observe, “behaving parametrically with respect to the market does not require building models of other agents.” Qua regularity of behaviour, a convention changes the situation – from the point of view of a given agent – from a strategic problem to a parametric problem. The agent no longer needs to worry about being taken advantage of if she cooperates. More cynically, of course, we can note that she also knows that she will automatically be successful in

---

5 That is, rational choice requires (ideally) that we know for each possible action what the associated outcome will be.
6 A ‘parametric’ problem is a problem of optimizing given a particular environment. As Elster (1979, p.18) puts it, in reasoning parametrically the agent “will believe that he is the only one whose behaviour is variable, and that all the others are parameters for his decision problem.”
7 See also Chapter IV of (Gauthier 1986).
8 Gibbard (1991) rightly points out the difficulty in characterizing a situation definitively as either ‘parametric’ or ‘strategic.’
9 In terms of the Payoffs in a Prisoner’s Dilemma (see Appendix 1) she doesn’t risk receiving the Sucker’s payoff (zero).
exploiting other agents’ constraint if she chooses to defect. Given the existence of a strongly entrenched convention (such as the queuing convention), an individual can expect utter regularity from others. That is, her fellow bank customers will constitute an ‘environment’ in a fairly strong, non-reactive sense, since if the convention is well enough entrenched, the others in the queue may not even be prepared to react to vanishingly rare cases of nonconformity. However, as Elster (1979, p.18) points out, reasoning parametrically may be “irrational from a collective point of view, even if individually each actor satisfies the criteria of rationality.” That is, if each of us thinks that her own choices are the only variables, she will fail to see her choices as affecting the choices other people make. This sort of reasoning is sure to prevent the establishment of mutually advantageous patterns, and is just as sure to lead quickly to the destabilization of those patterns if they do somehow get established.

Denzau and North distinguish between mental models, ideologies (essentially clusters of mental models), and institutions. Conventions in the sense in which we have been discussing them are part mental model, part institution. Like institutions, they are clearly social structures whose existence is independent of the beliefs or commitments of any individual. But to the extent to which a given individual knows that a particular convention holds sway in her society – to the extent to which she is sure that that convention actually constitutes a social regularity of behaviour – that convention serves as her mental model of other agents (or at least of those agents whom she believes will adhere to the same convention). That is, if she believes that convention to be pervasive, she will

\[10\] In the Prisoner’s Dilemma, for example, she will automatically receive the Temptation payoff (3) if she defects.
expect – with some degree of certainty – that those with whom she interacts will act accordingly, and will modify her own actions appropriately.

It may not be immediately clear that conventions share the salient features of descriptive mental models described above. Conventions do in fact have both of the simplifying characteristics attributed by Denzau and North (1994) to ideologies, and are thus “shared mental models” in the relevant sense. First, conventions are generalized in that, to the extent to which they are descriptive, they typically purport to apply to a wide range of agents. The queuing convention, for example, describes the behaviour of, at least, all North American bank customers. Thus to the extent to which they are accurately descriptive, they obviate the need to base assumptions about how other agents will act on knowledge of the particular characteristics of those agents. We all learn, at some point or another, what conventions are at play in our communities. Our knowledge of a pervasive convention, then, provides us with a mental model of a large number of individuals. Secondly, conventions also provide shared mental models in that, as historical entities, they can be learned from other agents. We typically learn from our parents which conventions are relevant in our communities; upon entering new communities, we usually try to find out from those better informed than us which conventions hold there. Such sharing obviates the need for each agent to determine (by lengthy observation and costly trial-and-error) both how others are likely to act and how she will in turn be expected to behave.

When the problem at hand involves collective action (and I argue in Chapter 4 below that moral constraint in general constitutes a collective action problem), there may be an even stronger incentive to rely on conventions or ideologies for one’s mental maps of other agents. In collective action situations, the effect which any given individual has on the
outcome of action is, *ceteris paribus*, inversely proportional to the number of agents involved. Downs (1957) has argued that the less an agent sees herself as being able to influence the outcome of a situation, the less motivated she will be to invest cognitive resources in the quality of her mental models.\(^{11}\) Accurately predicting how others will act only matters to the extent to which it guides our own decisions. Why should I expend energy on trying to predict how others will act if my actions don’t matter much in the first place? In activities requiring the input of a large number of agents, each agent will likely see her own actions as having little effect. So she will be unwilling to do much information gathering about her fellows, and thus is likely to rely on whatever conventions are known to exist to provide information about how others will act. We will move now to look at some examples.

Let us look first at a trivial coordination game. If you and I are in our cars, approaching each other on a narrow road, we need to decide on which side to try to pass each other. This is not something we can decide rationally: this symmetrical coordination game has two equilibria ([I go left, you go left] and [I go right, you go right]), and there is no *a priori* way to chose between them. But if each of us knows that there is a pervasive convention involving driving on the right, then each of us has a reliable mental model of the other. Given this mental model, each of us knows what to expect from the other, and can act accordingly.\(^{12}\) The fact that this model is both general and shared means that I can

---

\(^{11}\) Downs (1957) is cited in (Denzau and North 1994, p.11).

\(^{12}\) Sugden (1990) casts this in terms of coordinating around shared experiences.
expect it to help me reliably predict your behaviour, even despite my limited experience with driving and the fact that you are an absolute stranger.\textsuperscript{13}

Next, let us look at an example involving the convention of truth-telling. In many cases, our actions are guided by our beliefs about the veracity of other people’s utterances. Should I bother to take an umbrella when I leave the house, just because you tell me it’s going to rain? The fact that I know that a convention of truthfulness is generally well-established in our society means that I will assume (absent evidence to the contrary) that you speak the truth. And I will assume this about you even should you happen to be a stranger to me (though credence may be inversely proportional to the importance of the matter). This mental model that I have of you (and of others) also lets me know – crucially – that you are likely to respond badly should I mislead you.

Thus conventions provide crucial information about other agents, information that may critically affect the choices we make in social situations. Strategic decision making requires that we have some knowledge of how our fellow agents will act. Well-known conventions provide information both about how other agents are likely to act in the present moment, and about how they are likely to react after the fact to our decisions. Both of these kinds of information are crucial to agents wishing to choose rationally in social situations.

**Part 2: How Conventions Motivate Choice**

Next, let us look at the ways in which conventions motivate agents to select certain actions over others. Game theory provides a lens through which to look at the ways in

\textsuperscript{13} Brown (1995, p.18) refers to the rules of the road as helping us “overcome our fear of strangers hurtling toward us at sixty miles an hour.”
which conventions affect the choice situations with which agents are faced. Conventions as described here help us get along by providing solutions to problems of partial coordination. Thus conventions ameliorate situations that might otherwise have been worse for all concerned. But in game theoretic terms, problems don’t have solutions, they have *equilibria*. An example: in the absence of a convention (or other motivating factors such as sanctions), the strategic situation facing the bank customers is that of the Prisoner’s Dilemma:

![Figure 3.1: “Bank Customer’s Dilemma”](image)

The dominant strategy for each agent clearly is to jump the queue: no matter what everyone else does, any given agent does better to jump the queue. If others line up passively, then my unilateral aggression will get me quickly to the head of the queue. And if others push

---

14 A Nash equilibrium results when each of the players in a game has no incentive to change her strategy, i.e., when no agent can improve her lot by unilaterally changing her strategy. The rationality assumptions of game theory dictate that agents will always find equilibria where they exist.

15 For the general form of the Prisoner’s Dilemma, see Appendix 1. In any given instance the payoffs in the Bank Queue game will in fact be (unlike the payoffs for the Prisoner’s Dilemma) asymmetrical: whoever is ahead of me in line gains more than I do from our adherence to the convention. But in the long run, each of us will at some point benefit (i.e., be first in line), and so the game takes on the structure of an Iterated Prisoner’s Dilemma.
and shove, then I can only lose out by standing patiently where the queue would have been. No matter what others do, I should push. That is what game theory dictates a rational agent will do. In fact, since pushing is the rational choice for each, the [push, push] outcome will be the sole equilibrium in this situation. Yet that almost never happens in practice. Why? Why do seemingly rational people not find the equilibrium that rational choice theory says they should? This same question arises for all sorts of situations structured by conventions. At least four explanations are possible, from the point of view of traditional game theory. We will examine and reject these in turn.

One possibility would be to remove such situations from the domain of rational decision making altogether, by pointing out that the existence of a reliable convention may mean that the agent perceives no need to make a decision at all. Humans are great copiers and followers of others. When we walk into a bank and line up, we are simply doing as everyone else is doing, or what we have seen others do in the past. Game theory – a theory of decision-making – may then be excused for dealing badly with convention-following, since such behaviour does not seem to involve decision-making.

This argument is enlightening, but also perhaps misleading. It is enlightening in that it is indeed important to notice that situations governed by conventions may not be experienced as situations calling for choice. Most people, upon entering a bank, do not see

---

16 Note that the queuing convention is not universal, even within our culture: think of the news stories of shoppers pushing and shoving to get the latest fad Christmas toy. Think also of the Jobs board at the Eastern A.P.A. during lean years. It seems that the near anonymity of these situations, along with the fact that they approximate 'one-shot' games, hinders the development of what would be mutually advantageous conventions.

17 Evidence of the strength of the queuing convention in some places: (Hardin 1995, p.249) cites a New York Times article (19 December 1993, p.1.20) that described a bread line that stretched for three miles through Belgrade!
themselves as facing a decision about whether or not to join the queue. And that fact tells us something about the power and subtlety of conventions (and other soft institutions). But the argument is misleading in that it assumes that game-theoretic analysis applies only to situations involving conscious decisions. On the contrary, the move to apply game theory to non-deliberators is one of the great recent advances in the field. Further, it should be noted that copying the behaviour of other agents will sometimes, perhaps even often, be in fact a conscious rational decision. For a wide range of everyday activities, at least, copying what other people do is an effective, if conservative, strategy. To say that a given instance of convention-following amounts to copying what others do does not automatically imply that no conscious choice is made, and thus does not in itself remove conventions from the domain of rational choice theory.

This argument (i.e., that in queuing we make no decision at all) is misleading also in that it could only apply to conventions that are fully pervasive, and fully ingrained into the behavioural patterns of each individual. Only where behaviour is essentially automatic would it be prima facie fair to say that no rational choice is made. Actions that involve following conventions of which we are consciously aware, and which we evaluate on some level at least occasionally, clearly cannot be explained away in this manner. At least sometimes we do consciously evaluate our options and choose to act according to convention.

---

18 See Ross (1996). The classic reference in this regard is (Maynard Smith 1982)'s application of game theory to animal behaviour via evolutionary theory. Such extensions of game theory often require that the numbers in payoff matrices be taken to represent something (e.g., interests, offspring) other than subjective preferences.
A second possible explanation for bank customers’ ‘odd’ cooperative behaviour might be that they face not a one-shot game, but an iterated game. It is well known that, given repeated interaction, cooperation can be a stable equilibrium. That is, it can be rational to cooperate if one’s future opportunities depend upon the reputation one builds in the present. This is the wrong explanation for the success of the bank line-up convention, however. One of the conditions required for cooperation to become the equilibrium strategy in iterated games is that the same agents must meet each other again and again and remember each other’s past behaviour. Bank line-ups, on the other hand, are largely anonymous. Small towns aside, we usually do not line up with people we remember seeing before. That is the beauty of a true convention: it can involve strangers cooperating, despite not expecting to see each other again. Thus iteration alone cannot explain the force of conventions.

A third possible explanation would be to argue that some force or combination of forces is changing the payoffs of the game. The payoff numbers used by game theory are supposed to be descriptive of people’s preferences (usually ordinal rankings over outcomes), and preferences are taken to be represented in actions. Game theory rules out, by definition, counter-preferential action. Now by all appearances, people do not – contrary to the game matrix above – value the outcome associated with unilateral pushing over the outcome associated with multilateral queuing. If people do not in fact choose the “push” strategy in such situations, it must (according to traditional game theory) be because their preferences are not well represented by the existing payoff matrix. They may, for example,

---

19 This is the “Folk Theorem” that arises again in Chapter 4 below.
have a preference for order (over disorder), or a preference for avoiding other people’s disapproval.

So the existence of a reliable convention might be taken to alter behaviour by changing the payoffs associated with various outcomes. Indeed, the existence of a reliable convention can change payoffs in one of two ways. First, an agent who has internalized a convention may prefer actions which are in accordance with that convention. Skyrms (1996, p.28) claims that there is “no principled reason why norms of fairness cannot be reflected in [agents’] utilities.” Secondly, if most people (or a sufficient number) are willing to impose sanctions on those exhibiting non-cooperative behaviour, this effectively decreases the payoff associated with defection. Conformity may result in a variety of social rewards (which may include words of commendation, smiles, or approving glances) and psychological rewards (including perhaps feelings of fellowship and the knowledge that others are sure to approve of one’s actions). Given a broad willingness to impose social sanctions, the outcome matrix for the bank-customer’s situation – no longer a dilemma – looks like this:

<table>
<thead>
<tr>
<th></th>
<th>Line Up</th>
<th>Push</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line Up</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Push</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 3.2: “Bank Customer’s situation, given broad willingness to impose sanctions”

---

20 The existence of a convention can change the payoff structure in another way. As noted above, the regularity of behaviour implied by the existence of a convention can imply that the situation calls for parametric, rather than strategic choice.
Once we revise the payoffs this way, we see that lining up is in fact the equilibrium outcome, and theory coincides with observed practice. This tack is also the one advocated, for example, by leading game theorist Ken Binmore (1992, p.313). LaCasse and Ross (1998, p.348) use this approach to criticize Gauthier's claim that constrained maximization can let agents 'solve' the one-shot Prisoner’s Dilemma. They argue that if in a particular case temptation does not trump the appeal of coordination, then “that case is not an instance of the one-shot PD” (p.348). On this account, we are wrong to describe the members of a bank queue as facing anything like a Prisoner’s Dilemma.

However, this explanation of why people cooperate in such situations implies that this game is in fact embedded in a larger strategic situation, namely one that involves decisions about whether or not to impose sanctions. Thus it seems we should focus on the dynamics of that game. This may sometimes be an attractive move. Yet we should be cautious about this approach, for it is likely to obscure crucial facts. Namely, it obscures the important fact that there is a problem here to be solved. Revising the payoff matrix to reflect people’s actual behaviour hides the fact that something (namely, a convention) is doing some work here; it hides the fact that people do not simply prefer the cooperative outcome over the outcome associated with unilateral defection. That is, revising the payoff matrix as in Figure 3.2 seems to imply that people simply like lining up. David Schmitdz (1991, p.61) agrees that we should not revise payoff matrices so that they account for all motivational factors:

21 This approach is taken by (Heath, 1996).
22 These decisions can themselves be difficult, given the costs associated with imposing sanctions.
23 As (Gibbard 1990, p.802) suggests, equilibrium solutions have the advantages of “searching under the lamp.” That is, in general if interaction can be seen as being in equilibrium, we have a better chance of explaining it.
One could insist that a proper Prisoner’s Dilemma matrix must capture everything that the agents involved care about. Satisfying this requirement would have the advantage that the numbers depicted by the matrix could be taken to entail a preference ranking. I do not impose this requirement because I think it is impractical. It implies that we could never know that agents were actually in a Prisoner’s Dilemma unless we knew everything that might conceivably affect their preference rankings. We would have developed a conceptual apparatus for nothing.

Finally, we might explain the bank customers’ ‘odd’ queuing behaviour by suggesting that people generally (or bank customers in particular) are not always rational in the sense implied by traditional equilibrium analysis. Equilibrium analysis applies to rational agents.\(^\text{24}\) If equilibrium analysis does not work on bank customers, then they must be, in some sense, irrational.\(^\text{25}\) This explanation is clearly unsatisfying. The sorts of behaviour in question here are too common, and too useful, for this analysis to be attractive. We need not be zealous defenders of rationality in order to balk at calling a broad range of useful behaviours ‘irrational.’ Rather than calling them irrational, we might more reasonably note that those acceding to the queuing convention do not conform well to the homo economicus model of human motivation.

Thus none of the four explanations available to traditional game theory seems adequate to deal with the apparent divergence of convention followers from the predictions of rational choice theory. The actions involved in following conventions (other than purely coordinative conventions) are actions out of equilibrium. And this cannot be dealt with adequately by claiming that in following conventions we make no decisions at all, by

\(^{24}\) Or to optimising processes, in the case of evolutionary game theory. See (Maynard Smith 1982).

\(^{25}\) Binmore (1992) writes that “game theory...is mostly about what happens when people interact in a rational manner” (emphasis original). For Binmore, the “mostly” part signals only that game theory can apply to non-rational (but optimising) evolutionary processes, not that it can apply to quasi-rational agents.
pointing to the fact that conventions involve iterated games, by suggesting that the payoff matrices used fail to represent accurately agents' preferences, or by refusing to call convention followers rational. Next, I discuss a more radical alternative.

**Heath's Multidimensional Theory of Rational Action**

Rather than call 'irrational' any action that fails to maximize expected utilities, Joseph Heath (1997) suggests that we look beyond utility maximization in order to broaden the range of goals to which we can point in explaining action. Heath proposes that we look at what other sorts of factors besides outcomes can give agents reasons for action. Heath proposes a multidimensional theory of rational action. His account begins with the standard instrumental conception of rational action. He points out (as others have argued before\(^{26}\)) that the instrumental conception is limited by the indeterminacy of instrumental rationality in many social situations. In particular, instrumental rationality often underdetermines action in such situations, given that any agent’s choice may be conditional upon other agents’ choices, which can in turn be conditional, leading to a regress of anticipations.\(^{27}\) A theory of rational choice based solely upon the pursuit of desirable outcomes fails to determine action when the actions that other agents will take (and with which the agent’s own actions will interact to produce outcomes) is underdetermined by what we know about them. This points to the need for additional action-selection criteria. Heath’s solution is to

\(^{26}\) See, for example, Robert Sugden (1990).

\(^{27}\) This will be a problem in any game with multiple equilibria, such as either Battle of the Sexes or Chicken. In such games, what I should do depends on what you will do, which depends on what you think I will do, and so on.
postulate that actions themselves may be invested with significance aside from the significance attached to their consequences.

Heath (1997, p.28) proposes that agents can in fact have not just instrumental but also independent affective and normative reasons for action. The first of these three classes of reasons – instrumental reasons – is of course uncontroversial. Instrumental reasons are the motivating factors assumed by traditional rational choice theory. The last two – affective and normative reasons – represent different ways in which actions can be invested with significance that is non-instrumental. Let us examine affective reasons, and return to normative reasons later. As Heath (1997, p.20) notes, the “standard assumption is that agents only invest outcomes with motivational significance.” But Heath points out that there seems to be no good reason why only outcomes, and never actions, can be invested with special significance for agents. This point is surely right: everyday experience reveals that actions are often attractive despite their consequences. One may desire a cigarette, despite knowing that the immediate gratification is not worth (based on one’s own values) the long-term impacts on one’s health. Heath designates the criterion governing direct (non-consequential) selection of an action an urge. Since urges have as their object actions themselves, rather than consequences, they do not refer to other agents’ strategies, and hence cannot produce the aforementioned regress of anticipations. Using urges as a way of halting this regress is what Heath refers to as the ‘psychological solution’ to the looping (i.e., infinite regress) problem.

28 According to (Gibbard 1990, p.795), such non-consequential criteria, when publicly known, can also help solve the assurance problem: “Some emotional mechanisms give a person firmness of purpose by attaching a strong immediate feeling to actions that are out of keeping with that purpose.” (emphasis added)
Heath notes that affective reasons for action – urges – cannot be treated simply as an extension of the set of desires. That is, we cannot simply stipulate that agents have preferences for performing certain actions, and write those preferences into the relevant payoff matrices. This is surely the approach that would be suggested by traditional rational choice theory: the value placed on the action itself should modify the payoffs represented in the standard payoff matrix. Thus if an individual acts upon an urge to smoke, this would imply that she values immediate gratification (of this particular sort) over (some change in) her long-term health. But writing such preferences into the payoffs of the game results, according to (Heath 1997, p.22) “in nothing other than a net loss in information,” obscuring “the fact that some component of the agent’s motive for performing the action is non-instrumental....”

It seems that we now must have, then, a theory under which agents can have at least two different types of reasons for actions: instrumental reasons based on preferences, and affective reasons based on urges.

Heath (1997, p. 23) points out, however, that urges fall short as a solution to the equilibrium-selection problem if agents have divergent urges. Note that this problem stems not from the possibility that agents might have different urges. If we are playing Chicken (see Appendix 1), and if I know that your urge is to continue straight and you know that my urge is to swerve, I will swerve and you will go straight. Here, having different urges works out relatively well in terms of solving the equilibrium selection problem. But imagine a situation in which we are trying to coordinate on a meeting place without communicating.

Danielson (1995, p. 302) similarly warns us of “how easy it is to beg the question by building assumptions into the presentation of the situation.”

Of course, this equilibrium is better for you, since I lose face by swerving in Chicken, and you “win” by not swerving. But not finding this equilibrium could be much worse for both of us, if we both decide instead not to swerve.
If I know that your favourite spot in town is the Library (i.e., you have a standing urge to go there), then I have an instrumental reason to go to the Library (i.e., going to the library will help me achieve my goal of meeting with you). But if it is also the case that you know that my favourite spot is the mall, then you have an instrumental reason to go there. Here our different urges are a problem. What we need, then, according to Heath (1997, p.24), are reasons for action that are complementary or convergent. To fill this need, Heath (1997, p.24) points to the existence of norms which “...eliminate most of the guesswork involved with coordination through psychological propensities.” Norms, then, are in effect shared urges on Heath’s account. Heath (1997, p.25) calls a norm adopted as a selection criterion for actions a “principle.” Principles, then, constitute shared reasons for choosing certain actions – reasons that are generally independent of outcomes.

Note that principles seen in this way have a dual nature. On one hand, Heath’s account is consistent with Elster’s (1989a, p.98; 1989b, p.113) claim (discussed in Chapter 2 above) that norms are deontological in nature, rather than outcome-oriented. On the other hand, it is worth emphasizing the instrumental role attributed to norms; that is, the existence of at least some non-instrumental norms can be endorsed for consequentialist reasons. This might be seen as akin to a rule-utilitarian defence of rights, which gives a consequentialist explanation for the need to ignore consequences. This seems to reconcile

---

31 Notice that having the same urges can also be very bad, as when two of us have the same urge not to swerve when playing Chicken.

32 Whether the reasons need to be complementary rather than convergent will depend upon the situation. In the Prisoner’s Dilemma, the right sorts of convergent reasons will be helpful; in Chicken (as noted above) complementarity is required.

33 For a sophisticated defence of a utilitarian approach to rights, see chapters 3 and 4 of Hardin (1988).
the (at least partly) deontic *phenomenology* of moral norms with their functionalist *rationale.*

Heath is not alone in his contention that we should depart from traditional game theory in moving beyond the purely instrumental conception of motivation. David Schmidtz (1991), for example, argues that certain normative beliefs (e.g., that "squealing" is simply impermissible in the Prisoner’s Dilemma) rule out certain kinds of behaviour altogether. Such a factor does not make defecting any less attractive – any less preferred – than it would be otherwise. Rather it seems more appropriate to say that it "entirely eliminates [defecting] from your list of options" (Schmidtz 1991, p.60). To return to our earlier example, this departure from traditional game theory allows us to retain the standard Prisoner’s Dilemma diagram to illuminate the pre-conventional structure of the bank queue situation, and further specify that the existence of the queuing convention gives agents non-consequential principled reasons for not pushing.

The remaining problem with this account, according to Heath (1997, p.26), lies in explaining what to do when various sorts of reasons now on the agenda (i.e., instrumental reasons, urges, and principles) conflict. "The agent’s problem here is not with finding a particular justification, but with selecting a standard of justification" (Heath 1997, p.27). It should be pointed out, however, that this is a theoretical problem which may not be a problem in practice. For a give situation, instrumental reasons, urges, and principles may

---

34 In a similar vein, Harman (1983, p.123) argues that although once *accepted*, a moral rule acts to "restrict further appeals to self-interest," the point of the rule still has to do with mutual benefit, so that "self-interest never becomes completely irrelevant."

35 This loose approach to game theory is what allows Schmidtz (1991, p.76) to claim that certain kinds of contracts (what he calls "assurance contracts") make cooperation a dominant strategy in the Prisoner’s Dilemma. Traditional game theory can make no sense of such claims: either cooperation is a *dominated* strategy, or what we have is no longer a PD.
conflict in theory, but the agent may never experience that conflict as a problem. She may see these reasons as conflicting, but at the same time may be very certain which sort of reason is overriding. For example, she might realize that her instrumental reasons not to push to the front of the bank lineup (i.e., her fear of some sort of retribution), along with her principled reasons for not doing so (her ‘gut’ feeling that it is bad to do so), conflict with her desperate urge to go ahead and do so, and yet never question whether she should (or will) in fact do so. In other situations, the agent’s urges (e.g., to eat one more helping) may be so strong that it never occurs to her that other sorts of reasons (e.g., the disutility of an expanded waistline, or principles forbidding avarice) might apply.

There will, however, be times when normative principles and instrumental reasons are in fact experienced as in conflict. Heath argues that in most such cases, principles will likely take precedence. Heath (1997, p.29) postulates that “[s]uccessful socialization...has the effect of producing individuals who, in general, assign normative reasons priority over instrumental, and instrumental reasons priority over affective.” Thus well-socialized individuals choose morally rather than selfishly, and choose prudently rather than impulsively. Heath’s response here is plausible, but it is intended only as descriptive, not prescriptive. The remaining question of what ought to be done when normative principles and instrumental reasons do in fact conflict I set aside until Chapter 5, where I defend adherence to convention from the points of view of both morality and rationality.

Heath’s answer to our original question of how conventions affect choices should now be clear. Conventions are a subclass of the norms which Heath argues ought to be
regarded as direct, non-instrumental influences on our decisions. They provide reasons for acting – namely principles – that influence agents at the affective level.\(^{36}\)

Heath’s concern, again, is primarily with finding an account of rational action that avoids the indeterminacy problem. Heath’s (1997, p.31) argument is that a body of social norms ensures that “each social interaction problem can be given a determinate solution.” Norms can thus step in as supplementary motivators when instrumental rationality fails to ensure the selection of equilibria. If conventions are matters of contingency – that is, if any given convention is but one of many possible solutions to a problem – then we can say that the existence of a particular convention solves an equilibrium-selection problem. \textit{Given a desire to cooperate}, for example, in arriving at the bank, each agent in principle faces a problem of deciding just what action cooperation demands. Indeed, there are lots of ways\(^{37}\) of assigning priority to bank customers. Aside from queuing, other possible solutions include drawing straws, the game of Rock-Paper-Scissors, taking numbers, sorting by age, and so on. Without prior knowledge of local customs, two people entering a bank at the same time would have no way of deciding whether to form a queue or draw straws or what have you. As noted above, instrumental rationality in itself is insufficient, here. But given Heath’s model, we see how the existence of a norm (more specifically in this case, a convention) means that each agent knows what she is ‘supposed’ to do, and thus will allow the agents to coordinate upon one of the many plausible equilibria.

However, our conclusion about conventions can be stronger still. Heath argues that norms make up for one shortcoming of instrumental reasoning, namely its frequent

---

\(^{36}\) Note that in \textit{this} regard conventions are no different from prejudices.

\(^{37}\) Technically, an infinite number of ways, given an infinite number of criteria upon which we might think to base the assignment of priority.
indeterminacy. We can also argue, however, that the existence of shared norms can help make up for the other major shortcoming of instrumental reasoning, namely its tendency to drive agents to sub-optimal outcomes in certain social settings (e.g., in situations approximating the Prisoner’s Dilemma).

Note that the problem we began with - the bank customer’s dilemma - is not in fact strictly an equilibrium-selection problem. In the previous paragraph, we made it into an equilibrium-selection problem by assuming away conflict. That is, we assumed that the bank customers had already decided that they wanted to do the same thing. But in the bank-customer’s dilemma as described earlier, the problem was not one of equilibrium-selection; rather, the problem was that the equilibrium to which rational choice drives agents in such situations is a very bad one. We needed to account for the fact that real people - seemingly rational people - often do not find the equilibrium that rational choice theory says they should. Heath’s multidimensional theory of rational choice provides such an account. If rational agents seem sometimes to choose actions other than those dictated by instrumental rational choice theory, we can explain this in terms of the agents choosing instead based upon principles derived from conventions. Of course, a convention-based principle that dictates cooperative behaviour in a mixed-motive game must, if it is to be effective, necessarily be a stronger principle (i.e., more strongly prescriptive) than a principle that merely dictated which of several equally good equilibria to choose. That is, cooperation in a mixed motive game - such as the bank-customer’s dilemma - requires a convention with enough affective force to constitute for the agent a reason powerful enough to outweigh the instrumental reason which she has to defect.
In this Chapter, we have examined two ways in which conventions affect the choices which agents make. First, conventions have epistemic effects. Namely, they provide general, shared mental models of other agents that simplify the cognitive tasks associated with strategic interaction. Secondly, conventions also have motivational implications. Conventions can provide direct, non-instrumental reasons for performing certain actions, and thus can help us both to select among competing equilibria and to avoid mutually disadvantageous outcomes.
Chapter 4:

Conventions, Public Goods and Collective Action
This Chapter investigates the connection between conventions and collective action. If (as will be argued) moral constraints in general constitute public goods,¹ and if (as will also be argued) all (non-coercive) solutions to the collective action problems inherent in the provision of public goods are conventional, then moral constraints must be conventional. And if the establishing and maintaining of a social convention has the characteristics of a collective action problem, this suggests that one of the central problems of moral philosophy, namely the question of shared standards, can be enriched by looking at one of the central problems of political philosophy, namely the problem of collective action. Further, this implies that we should look to the literature on collective action in order to better understand the social dynamics at play in conventions.


I begin this chapter with a brief discussion of the concepts of "public goods" and "collective action." The problem of public goods and the worrisome logic of collective action have been important issues in political theory for the last two or three decades, but have received inadequate exploration by moral theorists.² According to David Schmidtz (1991, p.1) a public good "roughly speaking, is a good that can be produced only by collective action, but its production benefits people regardless of whether they join in the

¹ I am not the first to make this point. See also, for example, Danielson (1992, p.66).
² This is true largely, I suspect, due to the universalist prejudices so common among moral theorists.
collective effort.”

Schmidtz (1991) further characterises public goods as a subset of the larger class of collective goods. According to Schmidtz (1991, p.55) collective goods are goods the consumption of which is not subject to rivalry. That is, they are goods that any reasonable number of us may enjoy without diminishing each other’s enjoyment. Thus national defence, for example, is a collective good; the security I gain from the existence of a standing army (i.e., my ‘enjoyment’ of this product) in no way diminishes the security of my fellow citizens. (There are no worries about us each only getting a little piece of the pie.) The same goes for a litter-free sidewalk. In contrast, a house is clearly not a public good – if too many houseguests enjoy your space, you may begin to enjoy it less. Police protection is an example of an in-between case. Since the police cannot be everywhere at once, a given squad car patrolling your street is not patrolling mine.

Some collective goods are also public goods. Public goods, according to Schmidtz (1991, p.55), are collective goods that have as a further feature the fact that the

---

3 There are two other noteworthy ways in which public goods come into existence. First, a public good can come into existence through the efforts of a single person. I could, in principle, build a lighthouse – from which all would benefit – all by myself. More commonly public goods can be generated through the coercive power of governments. If the government is a democratic one, then we can give a Hobbesian account of the government itself (and derivatively all that it produces) as the product of collective action. But a despotic government, too, can produce public goods.

4 More controversially, professional standards might also be considered a public good in this technical sense. See Chapter 8 below.

5 David Braybrooke (1985, p.290) gives a more technical definition of a public good as a good that “is in joint supply, such that the marginal cost of supplying it, once in existence, to any additional consumer, is zero.”

6 That is, it costs no more for the army to protect 30 million Canadians than to protect 25 million Canadians, since the army protects our borders rather than protecting each of us as individuals.

7 Here, the public good is not the sidewalk (the enjoyment of which is diminished if too many people use it at once) but rather the cleanliness of the sidewalk.

8 Goods are public as a matter of degree; it is not a binary characteristic.
consumption of them is “nonexclusive.” That is, if such a good is made available at all, it will be available to all: excluding some people from benefiting will be either impossible or impractical. A clear example of this is a lighthouse: a lighthouse benefits all passing ships, regardless of whether a given ship’s owner contributed to the lighthouse’s construction. Environmental protection is another example: I benefit from others’ efforts at reducing emissions, regardless of whether I help, and regardless of whether others resent this.

Public goods are often so costly that their production requires the collective effort of a number of people. Very few individuals have the resources to build a lighthouse, or construct a highway or bridge, or clean up a riverbank. This being the case, public goods are famously subject to what has been called the ‘logic of collective action.’ The logic of collective action says roughly that difficulties arise in organising and motivating groups of people to produce public goods which, by definition, any individual will be able to enjoy regardless of whether she helps or not. When the good to be produced is a public good, I am better off sitting in the shade while my neighbours do all the work. I am tempted, then, (to use a public transit metaphor popular in game theory) to be a free rider. Since we are all in the same situation, my neighbours are all likely to think the same way I do, in which case none of us will contribute. Note that this failure to cooperate does not depend, however, on egoistic motivations; I may fail to contribute because I would rather spend my time helping elderly people at cross-walks, or take the money that I might have contributed to

---

9 As Gauthier (1986, p.87) puts it, the benefit derived by non-contributing ship-owners is a ‘positive externality.’

10 Some public goods are cheap: a backyard flower garden which I cannot prevent my neighbours from enjoying (visually) is a sort of public good. So is a roof-mounted weather-vane.

11 Thus as Gauthier (1986, p.88) puts it, public goods such as lighthouses “tend to be undersupplied by a market.”
our lighthouse project and give it instead to famine relief. The point is that my own preferences or values or moral beliefs or political values\textsuperscript{12} can prevent me from contributing to a particular group project.\textsuperscript{13} And even if some of us are more publicly minded, and think the group project is a top priority, there may not be enough of these virtuous souls to actually get the job done, so that those of us who do contribute will be wasting our time or energy or money. Thus even interested individuals may fail to contribute; even a real team player may hesitate to join in and lend a hand if she thinks her work will go to waste. This suggests that any project involving collective action faces serious hurdles. This is the pessimistic face of the logic of collective action.\textsuperscript{14}

Notice that to this point I have referred only to one-time projects. We get together to build lighthouse or we don’t. I contribute or I don’t. But what happens if we repeat the interaction? In one-off interactions involving public goods, there is no obvious incentive to contribute. Given the nature of public goods, I cannot readily be excluded from enjoying the fruits of your labours. But if you and I are going to interact again, I will need to take into account how my present actions will affect your willingness to cooperate with me in the future. And if it’s me worried about whether you’ll help out this time, I know that you know that if you don’t help, I won’t cooperate with you in the future. Given that we

\textsuperscript{12} Nozick (1974, p.94) gives the example of alternative political values – a preference that the group undertake a different collective action – as one reason why an individual might reasonably choose not to contribute.

\textsuperscript{13} Schmidtz (1991, p.83 ff.) uses the term “honest holdouts” to refer to people “who refuse to participate simply because they do not attach significant value to the good in question.”

\textsuperscript{14} The logic of collective action is for the most part also the logic of the n-person Prisoner’s Dilemma. Indeed, Russell Hardin (1982, p.25) has claimed that “the problem of collective action and the Prisoner’s dilemma are essentially the same.” I would argue that collective action problems are broader, and include more of the problems falling into the class of problems amenable to solution by convention than just the PD.
understand each other, and given that we all actually value the goods to be produced (or at least most of them), it becomes quite reasonable for each of us to contribute to the production of the public good. Indeed, it becomes flatly irrational of us not to.\(^\text{15}\) By making my future cooperation contingent upon your cooperation today, and vice versa, we can learn to cooperate. We have moved, now, from the pessimistic logic of collective action to reason for guarded optimism.

A further layer of complexity becomes apparent when we notice two important characteristics of attempts at punishing (by means of threats to withhold future cooperation from) those who fail to cooperate. First, notice that punishing defectors benefits everyone. If defectors are a general burden, then deterring defection is good for all. Secondly, punishment typically involves costs for whomever does the punishing.\(^\text{16}\) As Gibbard (1990, p.795) notes, “[i]f one punishes actively, one risks retaliation; if one punishes by cutting ties one forgoes chances for useful two-person cooperation.”\(^\text{17}\) Given these two characteristics, we see that punishment itself is a collective good, and thus punishing defectors is itself a collective action problem. Why should I bear the costs of punishing defectors? Why should I bear the costs of something that will benefit the whole group?

\(^\text{15}\) In technical terms, once you allow contingent strategies like Tit-For-Tat, the Iterated Prisoner’s dilemma can have coordination equilibria. This point is so commonly made that it has become known in the literature simply as “the Folk Theorem.” For a technical discussion, see Binmore (1992, p.357 ff.)

\(^\text{16}\) Sugden (1989, p.95) notes that given the human desire for approval, we can deter people from certain actions without actively punishing them, simply by disapproving of them. Sugden (p.95) refers to this as “a psychological externality: one person’s state of mind, as interpreted by another person, can affect that other person’s happiness or utility.” (emphasis original) But note that even this form of deterrence may not be cost-free. Bearing ill-will may impose a psychological burden on the bearer; further, it may make the offender less likely to engage in future mutually beneficial interaction.

\(^\text{17}\) Indeed, as (Danielson 1996) points out, in situations approximating the game of Chicken (see Appendix 1) resisting “defectors” is irrational even in one-off interactions.
Thus we lose the advantages gained by repeated interaction if we cannot figure out how to solve the collective action problem involved in imposing sanctions. Given the pessimistic logic of collective action, we might wonder why any projects requiring group input ever get accomplished.

Russell Hardin (1982, pp.173 ff.) points out that several characteristics of social interaction make the success of reciprocal strategies more plausible. We are all enmeshed in a range of repeated interactions in a range of areas. This has several implications. First, according to Hardin (p.174 ff.), we all have more opportunities to punish each other than would be the case if we were just involved in one repeated game. If a particular type of interaction is rare or unique, I will not fear your retribution if I fail to cooperate. But if we are neighbours, I know that we will interact in some way in the future, and will thus have an interest in maintaining good relations.

Second, Hardin observes (1982, p.180 ff.) that the fact that we are all involved in a range of interactions with each other also implies that we have lots of opportunities to learn about each other, either directly or by reputation. Information about each other – specifically how likely we are to cooperate and how likely we are to punish defection – is crucial in order for contingent strategies to produce good outcomes.\(^{18}\) Agents A and B may never have never worked together on a project like the one at hand. Thus neither of them knows if the other is trustworthy; nor does either of them know whether, or how harshly, the other is likely to react to a failure to shoulder a fair share of the burden. In such a situation, A and B may be unwilling to trust each other enough to get the project – which

\[\text{18} \text{ Such information is most readily available in dyadic interaction. As Hume (Treatise, III, II, VII. p. 538) put it, when two people interact, it is “easy for them to know each others mind…”} \]
they both agree is valuable — underway. But if the two of them have had some other type of interaction in the past (say, worked on some other sort of project) they may know enough about each other to be able to extrapolate to likely behaviour in the current situation. Similarly, if they have each worked with some third party, C, before, then C may be able to provide them some indication of each other’s tendencies. As economists Milgrom, North, and Weingast (1990) argue, “[e]ven in a community in which any particular pair of people meet rarely, it is still possible...for an individual’s reputation in the group as a whole to serve as a bond for his good and honest behavior toward each individual member.”

Finally, Hardin (1982, p.186-7) argues that the fact that we are all involved in a range of interactions with a range of people every day means that we can still make use of the strategic structures outlined above even if we seldom interact repeatedly with the same person on the same issue. That is, it does not matter that the interaction between two particular individuals is not part of a single iterated game: they can reap the benefits of iteration none the less. Hardin’s particular focus here is on situations approximating the Prisoner’s Dilemma. Hardin maintains (p.187) that if we looked at a set of diverse Prisoner’s Dilemma interactions involving a number of people, “the resulting ‘game’ would [ceteris paribus] be strategically equivalent to an iterated Prisoner’s Dilemma.” Thus some aspects of the complexity of social interaction may actually make collective action more feasible.

We return now to the idea of a ‘convention.’ We have noted that a convention in the sense in which we are interested is a regularity of social behaviour, sustained by a joint interest in coordination and an expectation that others will do their part. So whenever we have a sustained pattern of cooperative action toward a collective good, (and whenever that
pattern is not being sustained by a coercive central authority) we have a convention. In all repeated collective action problems, there are multiple possible conventions, and the right strategy for me depends very much on the local custom.

Part 2: Moral Constraints as Public Goods

Next I want to argue that moral constraints in general constitute public goods (though not ones produced by means of centralized authority), and hence pose collective action problems, and are hence conventional in nature.

I begin with an example of a moral rule that is quite easy to see as an instance of a collective action problem. In discussing collective action, I mentioned environmental protection projects as an example of a public good. It is not subject to rivalry: if the environment is saved for everyone else’s kids it is at the same time saved for my kids at no extra cost to anyone. And it is also non-exclusive: I benefit from the group’s attempts to preserve the environment, and I can’t be excluded from the good. But notice that this logic applies equally well to a broad range of issues in environmental ethics that have nothing to do with large tangible projects. The apparent moral prohibition on dumping motor oil down sewers is an example. Were this a one-shot game, it would be hard to see why I should cooperate. If everyone else cooperates (by disposing of their used oil in an ecologically sound manner), my bit of oil won’t matter much and I can avoid the significant inconvenience of proper disposal. If no one else at all cooperates, then the environment will be in serious trouble, and me going to the inconvenience of proper disposal won’t help anyway. Now consider what happens when we repeat the interaction. Once we repeat the interaction, my neighbours can tell me (or I can otherwise come to understand) that if I act
irresponsibly (i.e., fail to follow the convention), they'll punish me in some way or another during our future interactions. Notice that as an alternative convention, we could instead agree that it's alright if I dump oil, so long as I make up for that somehow, perhaps either by paying you money to compensate you for your share of the now-polluted water-table, or by doing some other good deed. My apparent obligation not to dump oil is now seen to be founded in a stable mutual expectation, or convention.

It is relatively easy to see environmental protection projects as collective action problems, and hence as conventional. But here we can play on the ambiguity of the word "environment" and point out that a vast range of activities have an impact on agents' environment. Many kinds of actions impact on agents' shared social environment. Any action that changes the nature of the social environment for the better (e.g., tends to promote stable property relations, or general non-violence, or punishment of wrong-doing) also constitutes a public good, and hence poses a collective action problem.

So let us move now to a less obvious example of moral constraint qua collective action problem: truth-telling. Why tell the truth? One might begin to reply by noting that communication becomes difficult if we don't generally tell the truth, and we usually see ourselves as having an interest in facilitating communication. This is true, but communication has more than one function. At the very least it can convey knowledge and it can entertain. Russell Hardin (1982, p.160) pictures three alternative standards with regard to veracity. One standard: tell me the truth. A second possible standard: tell me an interestingly embellished version of the truth. A third standard: don't tell me anything at all.

---

19 This idea of corporations internalizing externalities is famously discussed by Coase (1960). A proposal to let individuals engage in trade of rights to pollute can be found in Danielson (1993).
all. To that, we can add at least a fourth: tell me the truth most of the time, but lie to me if you think that will safeguard my feelings or interests. Which of these four standards is obligatory is surely a question of convention. The convention currently in place for whatever reasons is, it seems, the fourth one. In repeated interaction, people who either lied frequently or always told the truth regardless of the consequences would tend to face various sorts of social sanctions, ranging from verbal abuse and disapproving or hurt looks to avoidance of future contact. Note that here we may also see a variety of special conventions for specific domains. We may expect loved ones to bend the truth to preserve our feelings, expect nearly brutal truth from our academic supervisors, and accept with equanimity utter nonsense from those with whom we drink beer.

A third example will show that the provision of public ‘goods’ by convention need not constitute an optimal solution. Consider local customary variance from posted speed limits. For example, everybody in Ottawa knows that the de facto speed limit on Highway 417 is roughly 115 kph. Though the posted limit is 100, everybody knows (and some law enforcement officers will acknowledge) that no one would get a speeding ticket for driving 110. Further, drivers tend to get irritated when others only drive 100, and get worried when others drive much over 120. On Toronto’s Highway 401, however, speeds of 125 kph are not uncommon, and everyone knows that no one would get a speeding ticket for driving 120. Further, drivers who know the norms for both highways adjust their speed accordingly when they travel from one city to the other. We thus have local regularity. Further, such

---

20 Clearly the possible standards are limitless, given the option of having combinations of these four standards which would apply in a context-dependent way.
21 For a light-hearted discussion of the latter sort of convention, see Frankfurt (1988).
regularities constitute public goods. Everyone benefits from there being an agreed-upon 'speed limit.' According to Lave (1985, p.1163),

"... all current safety campaigns emphasize that 'speed kills.' They imply that the slower driver is the virtuous one and is helping protect himself and other drivers. It isn't so. To reduce fatalities, it is important that everyone drive at about the same speed. Thus the major consideration in choosing a speed limit is that it be obeyed. And the major consideration for police is to reduce variation, not speed, because slow drivers are as much a public hazard as fast ones."

Thus a shared driving speed benefits everyone. But drivers will have various reasons to deviate from the standard: some will be in a hurry and will thus exceed the conventional speed, while others will (erroneously) equate 'slow' with 'safe' and thus proceed at a more leisurely pace. And any given driver benefits from the safety that results from collective uniformity, regardless of whether she actually drives 'the speed of the traffic.' Adhering to this shared standard therefore poses a collective action problem.

This example is interesting in that the identification of moral constraint with the maintaining of a public good means that those who insist on driving the legal limit cannot claim the moral high ground. In this case, any form of deviation from convention can be seen to be unethical. Those Torontonians, for example, who insist on driving at the posted limit of 100 kph on the 401 (while everybody else is driving 120) imperil themselves and those around them. Thus moral constraint and legal requirements can conflict. As Schelling (1978, p.128) points out, "...the crucial element is often coordination. People need to do the right things at the right time in relation to what others are doing."

Let us look at one final example by means of a question: do graduate students or professors in the University of British Columbia's Philosophy Department have a moral obligation to go to Departmental Colloquia? Some members of the Department have
suggested as much. Others either don’t think so or knowingly act unethically on a regular basis (or are off adhering to other conventions in other domains). A lively colloquium series is clearly a public good: I benefit from my occasional attendance and from the prestige associated with a lively department, regardless of whether I contribute (by attending) on a regular basis. Such being the case, a colloquium series might be under-attended despite the fact that everyone wished the colloquium series were better attended. As we know from the discussion above, iterated collective action problems can be solved: important public goods can be generated. On the view expressed in this Thesis, I think that at least currently people do not have a clear obligation to attend colloquia. There is no stable pattern of expectations in place, and I can imagine a range of such patterns. At some universities, everybody attends every colloquium. Presumably in such a place I might have an obligation to attend. Other places don’t hold regular colloquia; presumably in such a place I would not have an obligation to attend. We see that the standard of behaviour in this domain, too, seems to be a matter of convention.

Having examined a number of examples, we can now generalize. Moral constraints in general constitute public goods, and hence pose collective action problems, and are hence conventional in nature. Thus the logic of (mixed-motive) conventions is the logic of collective action.

If moral constraints in general constitute public goods, and if all (non-coercive) solutions to the collective action problems inherent in the provision of public goods are conventional, then moral constraints must be conventional. We thus have an argument in favour of the central claim of this Thesis. Further, if the establishing and maintaining of a

---

22 Mixed-motive, as opposed to purely coordinative conventions which have a different logic.
social convention has the characteristics of a collective action problem, this suggests that we should look to the literature on collective action in order to better understand the social dynamics at play in conventions.
Chapter 5:
The Moral Significance of Social Conventions
This Chapter explores the possibility of a theory of normative ethics that places significant weight on conventions. We begin with a brief discussion of some worries about the motivational force of the dominant normative theories. We then move on to discuss the concept of a ‘convention’ and suggest that a convention-based theory goes some distance toward meeting these concerns. In the final section we will discuss the key weaknesses of a convention-based approach and suggest possible solutions. In the Chapters that follow, we will evaluate particular objections in detail: Chapter 6 will deal with problematic asymmetrical conventions, and Chapter 7 will deal with the apparent inability of a conventionalist account to generate moral obligations that cohere with our everyday moral intuitions.

Section 1: The Motivational Crisis in Mainstream Normative Theories

This Chapter is motivated in part by a serious weakness found in the dominant normative theories. David Gauthier (1991, p.15 ff.) rightly argues that utilitarianism and deontological theories fail to give a plausible explanation of why we ought to heed morality’s call. If morality involves constraints on choices and action, what reason could an agent have for accepting such constraints? This is particularly challenging in light of the fact that morality is typically seen as demanding of agents that they act, at least on occasion, in ways contrary to their own desires or interests. Gauthier argues that morality, as a justificatory framework, is jeopardized by the existence of an alternative framework, namely that of “deliberative justification.” This framework of deliberative justification –
what most rational choice theorists would refer to simply as ‘rationality’ – involves showing “that choices and actions maximize the agent’s expected utility, where utility is a measure of considered preferences” (Gauthier 1991, p.19). Simply put, this mode of justification is about agents attaining their own ends. Gauthier asks what attraction morality can possibly have, in light of this alternative justificatory framework.

A version of social contract theory may well, as Gauthier (1991, p.29) argues, pass this test by “exhibiting the rationality of our compliance with mutual, rationally agreed constraints....” That is, it can make sense to accept constraints that help the agent attain her goals. Gauthier provides convincing arguments to the effect that it can be rational for agents to adopt cooperative dispositions; i.e., that it can be rational for agents to adopt personality traits that make them trustworthy. Briefly stated, Gauthier’s contention is that agents with such dispositions will have profitable opportunities unavailable to less constrained agents.¹ He fails, however, to make a convincing argument for the rationality of cooperation itself. He argues (1991, p.25) that if it is rational to agree to something, then it must be rational to follow through on that agreement.² This argument has met with considerable skepticism.³ A more plausible line of argumentation might involve pointing out that if it is indeed rational to adopt certain dispositions, and if it is possible effectively to do so, then choices about whether to cooperate in particular situations just do not arise.⁴ For Gauthier is arguing not simply for the substantive rationality of acting like one has adopted

---

¹ That is, such agents can solve the “assurance problem.”
² Indeed, Gauthier is still trying to defend this thesis. See, for example, (Gauthier 1998).
³ See for example (Binmore 1994, p.80), (Heath 1996), (Skyrms 1996) and various essays in (Vallentyne 1991).
⁴ This leaves open, of course, the question of what psychological, social, or institutional mechanisms – if any – are capable of producing such dispositions.
a disposition to keep promises, but for the substantive rationality of actually adopting a disposition to keep promises. To the extent to which an agent actually is so disposed, there just is no question of compliance.

Gauthier’s account falls short in one other important area. He argues not merely for the rationality of constraint, but for the rationality of impartial constraint. In order to argue for impartial constraint, Gauthier (1991, p.27) is forced to appeal to the “instability” of any practice not meeting the standard of hypothetical consent. This argument is less than persuasive. Many practices not meeting the standard of impartiality seem to be quite stable, buoyed as they are by natural or social power differentials. Slavery, for example, seems to have been such a practice. The constraint involved in this practice is not impartial, and is thus not likely to be one that rational agents would agree to in advance, but history provides little evidence that it is unstable in any way that would worry – or motivate – particular slave owners.

For these reasons, we should accept only a narrower, more general version of Gauthier’s thesis, namely that it is rational – instrumentally useful – for agents to dispose themselves to at least some forms of counter-preferential constraint.⁵ In order to move closer to an argument for the rationality of compliance, I will depart from Gauthier’s account in that I will focus on boundedly-rational agents. That is, I will focus on agents differing from the “fully rational” agents assumed by rational choice theory in that they have limited cognitive capacities and limited time in which to make decisions. Hardin (1988, p.1) has argued that “[o]ur limited rationality and lack of information...are more fundamentally important than is generally recognized.” Hardin suggests that it has been a

---

⁵ Danielson (1992) arrives at a similar conclusion using the tools of computer modeling.
major failure of modern moral theory that it has focused on idealized agents, rather than on looking for moral systems that will work well for real, boundedly-rational humans.

I move now to propose a form of constraint that boundedly-rational agents should accept.

Section 2: Conventions

We have seen in earlier Chapters that social interaction – and thus moral choice – is shaped in important ways by a range of ‘soft institutions,’ including conventions, norms, practices, traditions, fads, etc. In this Dissertation, I have singled out conventions for examination. In the broadest sense, a convention is a regularity of behaviour that solves a coordination problem. It is a regularity in that it involves a group of people acting in a given way over a period of time. It ‘solves’ a coordination problem in the sense that agents who coordinate on a convention will do better than agents who fail to do so. A convention, then, is a social regularity, sustained by an interest in coordination and an expectation that others will do their part. But problems of pure coordination (for example, on which side of the road to drive, the meanings attributed to various words, etc.) are of little interest to ethics. Ethics has as its central subject matter constraints on action in social situations.

6 Note that saying that a convention constitutes a “solution” here does not imply that it is either a unique or an optimal solution. I return to the possibility of alternative conventions in Section 4 below.

7 Note here that the constraint is not merely on ‘self-interested’ or even ‘selfish’ action. Indeed, as Gauthier (1991, p.17) points out, ethics will often imply constraints on certain forms of beneficence. An example: most people would argue that there are moral limits on the things we may do to benefit our children.

8 We might be tempted to say that one can have moral obligations to oneself, e.g., obligations to look after one’s long-term interests instead of succumbing to rash impulses. We have other terms – “rationality” and “prudence” – for such constraints, however, so calling them “moral” erodes the usefulness of the term.
Situations of pure coordination require no real constraint. They only require constraint on caprice, in pursuit of the agent's immediate interests. For example, it is straightforwardly in my self-interest to follow the established convention of driving on the right in North America. I have no interest in deviating, and neither do I have any interest in you deviating. This situation is one of pure coordination, so all each of us cares about is that we do the same thing.

I am concerned instead with a large sub-class of conventions, namely those that provide solutions to problems involving partial coordination, or what game theorists call "mixed-motive" games. The most famous mixed-motive game is the Prisoner's Dilemma, which we have already discussed (see Appendix 1 for details). In general, a situation involves mixed motives in the relevant sense if and only if Agents A and B are in partial agreement with regard to the ordering of possible outcomes of their interaction. A convention (in the sense at issue here) is an accommodation between individuals facing such a situation. Such conventions are the sort with which Hume himself was concerned, particularly in Part II of Book III of his Treatise. Hume argued, for example, that both promise-keeping and "abstinence from the possessions of others" are conventions, that is, artifices devised to remedy natural conflict.

As a real-life example, let us return to a very common convention: the bank queue. If two of us walk into a bank at about the same time, and each needs access to the one

---

9 Situations approximating the Prisoner's Dilemma are seen by many social scientists as pervasive. Given the dilemma's title and its perverse logic, this apparent ubiquity might be shocking. But note that the PD has the same strategic structure as a simple exchange of goods. And as Hardin (1995, p.32) astutely notes, "Had the game originally been named 'exchange,' we would have expected it to be ubiquitous." (emphasis original)

10 I owe this example to Brown (1995, p.25 ff.)
available teller, we find ourselves in a situation in which our interests clearly conflict. We both prefer to be served first. Yet if we both actively pursue this preference by pushing and shoving, the confrontation may result in bloody noses, and neither of us will be served very soon. We are all better off (in the long run) if we (as a group) have a convention (like lining up) for settling peacefully our partial conflict of interests.\(^{11}\) The idealized 2-person version of this problem looks like this:

\[\begin{array}{c|c|c}
 & \text{Line up} & \text{Push} \\
\hline
\text{Line up} & \text{good} & \text{best} \\
\text{Push} & \text{worst} & \text{bad} \\
\end{array}\]

Figure 5.1: “Bank Customer’s Dilemma”

It should be clear that queuing is but one possible convention to meet the needs of this situation. It is conceivable that we could have had instead, for example, a convention that prioritized customers on the basis of age or urgency or Social Insurance Number, or a convention that had customers draw lots to determine priority.\(^{12}\) Thus queuing is not the one right answer, but it is our \textit{shared} answer.

A different example is provided by an arms race. For example, you and I are both better off if neither of us has nuclear weapons (less chance of a disastrous nuclear war, less money wasted on unused stockpiles). So we agree that a world without nukes is better than

\(^{11}\) On any given occasion, whoever is at the head of the line obviously benefits more from general adherence to the convention than does the person at the back of the line.

\(^{12}\) Of course, these alternatives are not all equally beneficial, but each of them seems better than no convention at all.
a world where we both have nukes. But each of us would prefer even more to be the only one with nuclear weapons (thereby reaping an enormous military advantage over the other). Thus we have a shared interest in reaching an arms accord, but each of us would in principle prefer to be able to deviate from it unilaterally. This situation has the strategic structure of a Prisoners' Dilemma. If we are able to achieve such an agreement (either explicit or tacit) in such a situation, it will count as a convention in the sense discussed here.

The general strategic structure (in idealized 2-person form) taken by situations capable of being solved by conventions is this:

![Figure 5.2: "The Strategic Structure of a Convention"]

If agents in such situations follow the convention (e.g., queuing at banks), they will all be better off than if they fail to do so (e.g., by pushing and shoving). Since each is better off under the convention, each has the above-mentioned "interest in coordination." Agents who unilaterally followed convention (i.e., unilaterally constrained their actions) would be worse off than they would be if they joined everyone else in ignoring the (would-be)

13 Indeed, the analysis of arms races as PDs is common (and historically important) in the literature. See (Gauthier 1986, p.80 ff.). However, the PD is not the only possible model for an arms race; see Measor (1983).
14 Note that the interest in coordination, here, is not the result of sanctions designed to enforce compliance.
convention. They thus only do whatever constitutes following the convention because of the above-mentioned "expectation that others will do their part." How well unilateral deviators do is indeterminate in the general case. There will be important classes of conventions in which unilateral defection makes agents worse-off (e.g., in the game, "Battle-of-the-Sexes"\textsuperscript{15}), and other important classes of conventions in which unilateral deviation makes agents better off (e.g., in the Prisoner's Dilemmas and games of Chicken).\textsuperscript{16}

Having explained what a convention is, I now move on to discuss their normative significance.

Section 3: Conventions as Appropriate Constraints

I expressed above my agreement with Gauthier's general argument in favour of the rationality of disposing oneself to some form of constraint. It seems clear that there will be constraining dispositions – moral attitudes or regularities of behaviour – that it is rational to adopt. But which dispositions or regularities? Candidates for such constraints might include those implied by various moral theories (utilitarianism, Rawlsian contractarianism, etc.) and those imposed by existing norms (altruism, patriarchy, the Golden Rule, etc.) and

\textsuperscript{15} In the Battle of the Sexes, 2 players prefer to coordinate their actions, but disagree over the relative merits of two Pareto-superior coordination points. (A situation is Pareto-superior to another if it is better for at least one person and no worse for anyone.) The game gets its name from a situation in which a man and a woman prefer spending the evening together to spending it apart, but disagree over whether to spend their time together at a football game or at the ballet. A more common example might involve an bilingual Anglophone and a bilingual Francophone deciding in whose mother tongue they should converse. Another example would be a situation in which two manufacturers want to make their computer components compatible, but disagree over whose proprietary technology to use to accomplish the feat. For the payoff matrix for this game, see Appendix 1.

\textsuperscript{16} See Appendix 1 for the matrices for these games.
conventions. I next propose that conventions constitute a class of soft social institutions meriting serious consideration as embodying such reasonable constraints.

The motivational question that normative ethics faces is complex. A normative theory faces two separate challenges when it comes to motivating the agent. First, as Gauthier points out, such a theory needs to be able to say why the individual should constrain the pursuit of her own interests at all. Secondly, we do not want simply to advocate constraint in all its forms: some forms are trivial, and some are oppressive. We need to be able to say which kinds of constraints agents ought to accept as appropriate moral constraints. I argue that conventions constitute at least a rough answer to the latter question: agents should constrain themselves in the ways demanded by adherence to established conventions (or to some sub-set thereof).

Prescriptions – constraints – come in varying degrees. Consider the different levels of force that must be attributed to a prescription in order for it to be motivationally effective in the following range of situations.

1) Prescriptions to cooperate in problems of pure coordination require minimal (indeed, zero) constraint,\(^\text{17}\) since agents gain nothing from unilateral deviation. For example, calling a truck a “truck” while in North America and a “lorry” while in England is straightforwardly in my self-interest. Usually, clear communication is one of my goals, and so cooperating in this way does not imply a constraint upon my pursuit of self-interest.

2) In a 2-player iterated Prisoner’s Dilemma, acceding to the prescription, “Cooperate,” is contrary to ‘modular rationality’ (requires action out-of-equilibrium, contrary to the individual’s immediate interests), and so requires some – but not very much

\(^{17}\text{...except for the above-noted constraint on pure capriciousness.}\)
constraint. If I know that your willingness to cooperate with me tomorrow depends on how I treat you today, then I must at least (or “only”) constrain myself to forgo immediate gratification (i.e., the benefits of unilateral defection) in favour of my own longer-term self-interest.

3) Cooperating in a one-shot problem of partial coordination, e.g., any one-shot PD (such as the decision whether to help a neighbour bring in her crops, given the knowledge that one of you is moving away after this season\textsuperscript{18}) is, \textit{ceteris paribus}, contrary to the individual’s interests. It involves action out-of-equilibrium, in the pursuit of benefits (e.g., a good reputation among those with whom one \textit{might} interact in the future) that may be very uncertain, and so requires significant constraint.

4) Making threats (to make both worse off than the equilibrium outcome) in defence of fairness (e.g., in the Ultimatum game, or in a bank lineup) typically involves accepting costs that far outweigh one’s share of the collective benefit, and only garners the individual uncertain reputational benefits. Doing so typically requires still more constraint.\textsuperscript{19} Constraint of this sort constitutes a difficult collective action problem, as discussed in Chapter 4 above.

\textsuperscript{18} The benchmark discussion of this example is to be found in Hume’s discussion of the conventional nature of promise-keeping. See his (\textit{Treatise}, III, II, V, p.520). See also (Gauthier 1991, pp.24-25.).

\textsuperscript{19} Actually, whether situation #3 or #4 involves more constraint will depend on the stakes in particular situations. Some threats, for example, will not be very costly to make or carry out. Nothing rides on the ordering of these two examples; they are for illustrative purposes only.
5) Finally, acts of pure beneficence (such as giving to famine relief) are strictly contrary to the individual's interests (assuming an absence of affective ties), and so may require relatively strenuous constraint.20

Situations of the first type – pure coordination problems – tend to be governed by the simplest21 of conventions, namely purely coordinative conventions. Given the existence of a convention covering such a situation, no one has any real interest either in failing to follow the convention, or in anyone else failing to follow the convention. Situations of the second, third and fourth types, embodying as they do problems of partial coordination, could also be subject to a convention, and such a convention would be just the type with which this Thesis is concerned. Since the third and fourth types of situation require more constraint than the second, a convention covering such situations would have to be strongly ingrained or enforced by social sanctions in order to be effective.22 Situations of the fifth type are, by definition, never governed by conventions (though they are of course often governed by norms of other kinds), given that situations involving pure disinterested beneficence involve no convergence of interests among the agents involved.23

A normative theory that identifies conventions as suitable moral constraints therefore draws a line: it would give moral weight to constraint in situations of types one24

---

20 This may seem counter-intuitive: many of us find such constraint easy. But we should ask what social and psychological forces are at play that make such constraint easy.
21 Such conventions are "simple" motivationally. Measured other ways, of course, some coordination conventions (e.g., at least some linguistic conventions) are extremely complex.
22 And of course, the imposing of sanctions constitutes a further problem of type Four.
23 Beneficence may, however, require coordination among benefactors. I set aside until Chapter 6 this question of higher-level conventions.
24 Whether constraint in Type One situations should be termed "moral" is debatable. If we take problems of pure coordination as too simple to require ethics (because they require no real constraint), it would perhaps be more appropriate to exclude conventions governing such situations from the class of moral conventions. However, it seems reasonable to count at least some purely coordinative conventions (such as "drive on the right") as moral conventions, in that they are
through four, but not to constraint in situations of type five. Whether or not we find this dividing line immediately plausible, we should count it as a virtue that such a theory would in fact provide such a line to distinguish reasonable constraints from supererogatory constraints. In this regard, recall that Gauthier's argument is in favour of stronger impartial constraint. On Gauthier's (1991) account, equally rational agents would demand of each other impartial constraint. Since rational agents would not accept - or expect others to accept - a disproportionately small gain from the cooperative endeavour, impartial constraint may require a renegotiation of the existing moral order, so that each individual gets as much as she could have bargained for from a fair initial bargaining position. Acceding to such a renegotiating would likely require a well-off individual to undertake acts that amount to one-sided beneficence in order to arrive at a social distribution consistent with what rational agents would agree to. Thus although Gauthier's appeal for impartiality is rooted in reciprocity, given present-day inequities, constraint of the kind Gauthier envisages would imply a duty of pure beneficence toward those less well-off. (This is odd, given Gauthier's (1986, p.225) claim that "redistribution cannot be part of a rational system of cooperation.") Of course, some degree of redistribution fits well with most people's moral intuitions. While Gauthier's bargaining solution puts limits on this

---

important enough that we would demand that people adhere to them even if (however unlikely) they wished to deviate. Further, as LaCasse and Ross (1998, p.342) point out, "there is no guarantee that players in a co-ordination game will in fact co-ordinate on the optimal outcome." An equilibrium that is hard to find (or easy to lose) might need to be moralized.

25 Gauthier (1986, p.6) writes that "Our theory must generate...constraints on the pursuit of individual interest or advantage that, being impartial, satisfy the traditional understanding of morality."

26 The details of Gauthier's bargaining solution can be found in (Gauthier 1986, Ch.V). Roughly, Gauthier's joint bargaining solution involves making the benefits of cooperation proportionate to the initial bargaining endowments - i.e., contribution - of the individuals involved.
redistributive argument (e.g., it stops short of demanding equality of condition), its
demands could be very great none the less.\textsuperscript{27}

Conventions have two other characteristics that make adherence to them a credible
form of constraint. First, it is clearly an advantage of conventions, where they exist, that
they constitute a \textit{shared} understanding of both how people \textit{do} act and how they \textit{should} act.
How agents do in fact generally act in a given sort of situation is important, given that a
particular convention is typically one among many alternatives. The existence of an
established convention serves as a salient point around which to coordinate.\textsuperscript{28} In many
commercial domains, for example, there is a convention involving a relaxed standard of
truth-telling. If A is attempting to sell her house to B, neither is likely to expect \textit{complete}
honesty from the other with regard to the negotiations over price. It is commonly
understood that a lack of candour is permissible in such situations. Thus A knows to take
with a grain of salt B’s claim that his current offer is his last. In other matters (e.g., whether
the house is structurally sound), we do expect complete honesty. Thus if A, in negotiating
with B, wants to know how seriously to take B’s utterances, it will help if she actually
knows what the relevant convention is. Similarly, good-natured B may not want to bluff if
she knows that A expects candour. If we want to coordinate our actions, we need to know
what others are doing. And a shared understanding of how people \textit{should} act is crucial,
given the pragmatic need which agents have to justify (i.e., to make reasonable) their
actions to their peers. I return to the question of justification below.

\textsuperscript{27} Gauthier (1986, p.105) uses a similar argument against the redistributive implications of
utilitarianism: “The utilitarian, in redistributing goods, necessarily imposes costs on some persons
in addition to those that have fully paid for the benefits they received in free interaction, and confers
benefits on others who have not paid their costs.”

\textsuperscript{28} For more on this, see (Sugden 1989).
Thus conventions serve the realistic needs of agents better than, for example, abstract aspirational norms or principles would. Aspirational norms will often be very bad coordination points, since it is difficult to know when they are shared. Thus even if A thinks that the world would be a better place if negotiations were carried out candidly, this simple conviction will serve her poorly as an indicator of how B will actually act. Aspirational norms also fail to meet the agent’s need to justify her actions to her peers if those norms are not shared. If the extant convention is one of buying alternating rounds of beer, I will be poorly served – i.e., I will get along poorly with my fellows – by acting (on the basis of an unshared principle) as if each of our contributions to our collective bar tab should be proportional to individual income.

Finally, conventions have the advantage of being, by definition, beneficial to all to whom they apply. That is, in certain situations members of a group with a convention will do better than a group in a parallel situation without a convention. Thus conventions embody both individual benefit and (at least rough) reciprocity. This pairing of characteristics should be expected to have a significant stabilizing effect. In comparison, a norm from adherence to which some individuals derived no benefit (or less benefit) is unlikely to command lasting allegiance without significant enforcement. A norm which commanded unilateral constraint would face similar concerns. My point here amounts to a more cautious version of a similar claim made by Gauthier. As noted in Section 1 above, Gauthier appeals to the ‘instability’ of any practice not meeting the standard of hypothetical consent. While I register a worry here about practices which embody neither mutual benefit

---

29 See Hardin (1995) for examples of strongly entrenched norms that offer no apparent benefit. But as Hardin suggests, such norms may provide (or have their genesis in the provision of) benefits to the group.
nor reciprocity, I do not go so far as Gauthier in asserting that this worry is enough to make anyone's support of such a practice *irrational*.

Conventions can thus be endorsed by boundedly-rational agents. Unlike ideally rational agents, boundedly-rational agents lack perfect knowledge, and thus need signals as to how others will act. Further, the outcomes generated by conventions are better (for all) than the sub-optimal outcomes to which the ideally rational agents of rational choice theory can be driven in similar situations. It is rational to employ whatever measures are available (either psychological or social) to dispose ourselves to comply with conventions.

**Section 4: Some Reservations**

The most obvious charge against any theory that attributes prescriptive weight to conventions is that it is overly conservative. If existing patterns of behaviour have moral force, how can we justify any action that seeks to change those patterns? Recall the bank scenario described earlier in this Chapter. The easy conformity that we find in the convention of lining up was offered as a tidy solution. Yet such rule-following could also be seen as mindless obedience. Why not look for a *better* solution? For one thing, the convention currently in place may be outdated. As Mackie (1977, p.121) points out, "the device of morality is beneficial because of certain contingent features of the human condition. But if they are contingent they may also have changed." For example, the old Scottish highland convention of offering hospitality to any and all who knocked on your
door is likely unworkable – indeed, dangerous – in modern urban environments. Conventions, then, can outlive their usefulness.

Further, although I have given no specific account of the origins of conventions, it seems likely that most of them have their origins in some blind social evolutionary process. There is no reason to expect such processes automatically to produce solutions that are optimal, though they may of course do so on occasion. As Hardin (1988, p.16) puts it, it would be an “odd epistemology that dictates that socially derived knowledge must always be superior to individually deduced knowledge....”31 Robert Sugden (1989, p.93) expresses this idea in the language of evolutionary game theory, noting that “conventions can be evolutionarily stable even if they are not Pareto-efficient.” Recall for example the idea, discussed in Chapter 4, of a conventional local speed limit. We pointed out that a shared standard of traffic speed is good for all; within reason, excess variation can be more dangerous than excess speed. But at some point, absolute speed must matter. What if the local convention involves speeds so high that they tax the average driver’s skills or perhaps induce ulcers? Since we can never be sure that any existing convention is the very best one for the situation, it seems clear that there are limits to the normative force we would want to accord to existing conventions.

Naturally, the fact that a convention is (by definition) useful is no reason not to question it. However, in pointing out that conventions constitute a useful class of

---

30 Smith (1776/1998, p.261) notes this tradition of hospitality, and points out that it is “common in all nations to whom commerce and manufactures are little known.” This suggests that it is a convention that meets the needs of a particular type of setting.

31 Emphasis in original. Hardin was referring in particular to the epistemology he attributes to G.E. Moore.
constraints, we are thereby warning against unthinking change. To dismiss a conventional standard as "mere habit" is to underestimate the work that conventions do. In many cases it will be both imprudent and irresponsible to destabilize an existing convention. Granted, it will sometimes be a very good thing that a given convention is destabilized or even toppled. But just as clearly, destabilizing one convention without replacing it with another could, in some domains, have disastrous effects. Think, for example, of the convention (discussed in Chapter 1) forbidding the killing of non-combatants during wartime, and the dreadful effects that the destabilization of that convention could have. In many situations, any of a number of fairly bad conventions may be preferable to having no convention at all. For example, the limited warfare that results from prohibiting the killing of non-combatants still exacts a terrible toll; thus we might reasonably say that it is a fairly bad convention. A convention of respecting the outcome of combat of single champions would in principle be better for all concerned, if it could be instituted and adhered to. But limited warfare is still better than a convention-less unlimited warfare of population against population.

Russell Hardin (1988, p.17-18) presents a further argument in favour of existing social rules. In particular, he defends the sort of rule utilitarianism that emphasizes common-sense moral rules, but his argument applies equally well to conventions. Hardin suggests that general limits on both reason and theory warrant skepticism regarding our

---

32 Brown (1995, p.19) compares a convention to "an old suit of clothes that fits well, although the wearing has made it threadbare. I am reluctant to throw it out until I have no other choice."

33 Hardin (1988, p.17) notes that a rule that helps us coordinate in mutually beneficial ways "might be inferior to others we could have if we were able to legislate a new rule, but without ability to legislate and change the rule it might still be better than not having a rule at all."

34 Or perhaps we should say the effect that this destabilization has already had. Witness the wartime atrocities committed in recent years in both Rwanda and the former Yugoslavia. Ignatieff (1998, 156 ff.) attributes this degradation of the rules of war to the de-professionalization of warfare.
ability to engineer new rules that are better than old ones. Hardin concludes (p.17) that “[s]ocial testing is often therefore a better guide for our action than is abstract deductive reasoning.” Thus where an existing rule is working even tolerably well, we should question whether we have the tools with which to do better.

We should also note that a focus on conventions is much less conservative than, for example, an account that attributes moral force to the broader class of norms (where norms are merely social regularities of behaviour that the individual takes as prescriptive). Norms as such need not advance the interests of any of the agents governed by them. Norms can dictate behaviour that is meaningless, or even very costly, to the agent. To grant moral force to norms as such would be the worst kind of conservatism, entrenching established patterns of behaviour regardless of their consequences. The conservatism of a convention-based approach is indeed moderate in comparison, since it endorses existing patterns of behaviour only when they are mutually beneficial. However, if we are to reach any general conclusions as to how much prescriptive force to accord to various conventions, we will require a subtler typology of conventions, the situations they govern, and the social mechanisms that enforce them.

In granting conventions prescriptive force, we must also acknowledge the possibility of pernicious conventions. Some conventions will have deleterious effects on persons not party to the convention. Indeed, there can be conventions aimed specifically at

35 Note that by “conservative,” here, is meant roughly ‘accepting of existing standards.’ If instead by “conservative” we meant ‘more restricted’ or ‘accepting of fewer possibilities,’ then conventions would indeed be more conservative than other norms, in that they must meet an additional criterion (i.e., that of mutual benefit).
36 Hardin (1995) discusses a range of norms to which individuals have been subject and which impose real burdens.
harming some third party. A very important modern example is the convention which protects those who make pirated copies of musical recordings and software for friends and family. Such activity is clearly illegal, and results in considerable lost revenue for the recording and software industries. Many people probably don’t know that such copying is illegal: but even among those who know that it is illegal, there is a conventional understanding that it is morally permissible. Indeed, some think it obligatory, and would think one either sanctimonious or ill-mannered if one refused to make a copy of one’s newly-purchased CD for a friend or relative. This convention benefits some at the expense of others. Other examples of pernicious conventions might include conventions of honour among thieves, conventions around contributing to war efforts, and the code of silence attributed to the members of various organisations.\footnote{With regard to the latter, both organized crime families and police unions come to mind.} Can we be comfortable with according the Mafia Code of Silence full normative force?

This discomfort with pernicious in-group moralities is one of the reasons why defenders of moral objectivism worry about conventionalism. Objectivists such as Louis Pojman, for example, contend that to give moral weight to contingent conventions and thus to step away from objectivity is to step onto a very slippery slope. If large social groups get to make up their own moral rules, they ask, why cannot smaller groups or even dyads do so as well? But this attempt at \textit{reductio ad absurdum} fails because it makes the illegitimate – and implausible – assumption that there is only one relevant framework for justification.

Pojman (1998, p.44) argues that the conventionalist perspective “permits almost any principle at all to count as moral.” Of course, Pojman does not believe this literally to be true. There may be limits on what even a conventionalist would admit as a \textit{moral} rule. By
Pojman’s own stipulation (p.41), morality has to do with “the amelioration of the human predicament,” and has “the minimal aim of preventing a Hobbesian state of nature.” The conventionalist can agree with – and indeed I will insist on – some (more individualistic) version of this proviso. Given this qualifier, Pojman’s claim is roughly right: conventionalism does seem to permit the standards promulgated by small groups to count as moral. But it is odd that he thinks that this is a failing. It would be both strange and tragic if only very large groups were capable of having moral rules. One wonders how far back up the slippery slope one must climb: what critical mass must a social group reach before it can have moral rules?\footnote{Nothing we have said so far suggests that a convention requires more than two people. It certainly seems plausible, for example, to imagine a range of conventions developing between the two inhabitants of a desert island. It is hard to see why such a system of conventions – especially if adhered to and internalized – should not be considered a moral system.}

Ethics – moral theory – is about what actions we can justify, and to whom we can justify them. The need to justify our actions to a larger social unit is what prevents us from saying that two students could form a club and “decide” that cheating on university tests is morally permissible. (This example is the absurdum to which (Pojman 1998) drives his reductio.) Other things being equal, the two of us are free to agree between ourselves on the terms of our interaction. In particular, we will be free to decide the terms of our interaction just so long as our interaction does not affect others.\footnote{My reply here is the same whether Pojman is concerned with the absolute size of a group or only with the relative size of sub-groups.}\footnote{See Hardin (1988, p.84 ff.) for an interesting analysis of rights in dyadic interaction (e.g., minimum wage laws) as a way of protecting persons outside of the dyad.}

Here I am not simply paying homage
to the liberal idea that each of us should have a range of unfettered action, a right to swing one's arm so long as one's swing falls short of one's neighbour's nose. I am pointing here rather to a fundamental point about the nature of justification. Justification is not an abstract quality: it is always a matter of justification to someone.\textsuperscript{40} So it is critical to distinguish between justification within a group and justification to outsiders. I see no \textit{prima facie} reason why we should think it absurd that a mafioso sees himself as having moral obligations that, for example, entail criminal activity.\textsuperscript{41} Further, I see no reason why, given this admission, we would be justified in asking whether this person \textit{really} had a moral obligation. In a very real sense he really \textit{would} have a moral obligation, relative to the moral code of one relevant group. Appeal to the moral code of the Mafia will work perfectly well for the task of justifying the mafioso's actions to other members of that group.

Of course this perceived obligation is unlikely to be seen as exculpatory by the larger social group. Indeed, why should it? The role (and content) of the Mafia's moral code can be explained in terms of the need to mediate interpersonal conflict (or lack of coordination) within that group. For example, it may be that the harsh code of justice according to which the Mafia deals with insiders enhances unity in mutually advantageous ways. That is, the Mafia's moral code is a set of conventions that serves the Mafia's

\textsuperscript{40} As Rawls (1971, p.580) writes, "justification is argument addressed to those who disagree with us." (Although, as Peter Danielson has pointed out to me, there is no reason justification must take the form of \textit{argument}. Justification at its simplest is merely the process – whatever its nature – of getting someone to think something justified.) For a discussion of the pervasiveness of this approach to justification as a methodology in contemporary political thought, see Norman (1998).

\textsuperscript{41} The parallel here with killing an enemy soldier is probably obvious. Though the parallel is not a clean one, I suspect that it is defensible. Any argument to the contrary would likely involve problematic theoretical claims about the obligations one 'acquires' by being born into a civil society.
interests (just as “our” moral code serves our interests). But where the interests of the Mafia conflict with the interests of the larger society, a framework intended to serve the Mafia’s interests alone will have no justificatory force at all in the eyes of that larger society. Thus we need not fall into the silly sort of relativism that says that anything goes, or that cultural respect implies cultural acceptance. But our previous assumption of the idea that all value is relative to the agent prevents us from assuming that there is one frame of evaluation that is in some sense the real source of justification. The search for a broader mode of justification is attractive, but seems undermotivated in those cases in which the broader framework includes individuals (or groups) about whose opinions the agent does not care, or to whose interests the agent is indeed antagonistic. The breadth of our frame of justification should be determined on pragmatic grounds. Thus the fact that conventions can be pernicious to outsiders – and thus cannot be justified to them – is lamentable, but is no objection to a convention-based theory.

In this regard, it should also be noted that a convention-based model of ethics would also have the general virtue of making transparent the limits of moral justification, and the particular virtue of making transparent the limits of justification of pernicious norms or conventions. A conventionalist approach of the kind outlined here makes clear that pernicious social conditions can never be justified simpliciter; they likely can only be justified to those who benefit from them. If, as seems likely, a conventionalist approach can

---

42 Naturally, to the extent to which the Mafia needs to cooperate with the rest of society, it will need to share some of the larger society’s conventions.

43 Similarly, Harman (1983, p.114) claims that to say that all morality consists in conventions “is not to say social arrangements are just whenever they are in accordance with the principles of justice accepted in that society. We can use our own principles of justice in judging the institutions of another society...."
to some extent endorse patterns of behaviour which benefit some at the expense of others, it at least has the merit of being unable to justify such patterns in an *absolute* way. If it can grant moral weight to existing patterns of the distribution of basic goods, at least conventionalism – unlike, for example, Nozickian notions of justice\(^{44}\) – makes no pretensions that such distributions are “really just” in some absolute sense. If it can give some endorsement to in-group moralities, at least such endorsement is drastically limited by the fact that the form of justification at play here is *explicitly* a narrow one. If value is subjective and relative, then justification can only be based on the interests (broadly defined) of the individual; appeal to Mafia conventions may explain, but never justify (to the larger society), the mafioso’s actions.\(^{45}\)

One further approach should be mentioned that might be used to address any remaining concerns about conservatism or about pernicious conventions. It might be possible to *supplement* a system of conventional constraints in ways that deal with the objections discussed above. It is likely that the source of these objections lies in the attempt to base obligation entirely on the rationality of constraint. It is just not plausible that it will always be *rational* to constrain oneself, or to dispose oneself to show constraint, in all the situations to which our existing moral intuitions apply. To think that a rich moral code that coheres with our natural sympathies can be generated from instrumental rationality is unrealistic. I am less dedicated to the attempt than Gauthier is, and thus am ready to admit that the set(s) of moral obligations generated as a result are in some sense unsatisfactory. We should think of these worries as pointing out the limits of convention, rather than

\(^{44}\) See Part II of (Nozick 1974).

\(^{45}\) In a similar vein, Harman (1983, 128) suggests that making explicit the self-interested motivation of morality would “lead us to greater clarity and honesty and less hypocrisy in moral argument.”
casting doubt on this project as a whole. It may be possible to supplement convention-based constraint among rational agents by adding an extra premise, in the form of an expanded theory of value. But we should strive -- for methodological reasons, if for no other reason -- to make that additional premise as weak as possible. For Kantian contractarians such as Rawls, that premise is the inherent equality -- equal moral worth -- of all moral agents.\textsuperscript{46} The arguments against this strong premise are too complex to discuss here. But I propose instead that we supplement our theory of conventions with a \textit{thin} value theory.\textsuperscript{47}

The value theory assumed so far (and discussed in Chapter 1) has been what is likely the thinnest of all, namely one (shared with Gauthier) that holds that value is both subjective and agent-relative. It may be possible to add to that a claim (not an argument, but a premise) about \textit{what things are valuable}. Strictly speaking, we should side with Gauthier (1986, esp. p.50-55) (and many others) in holding that value must be taken as relative. That is, things can only be valuable relative to some evaluative framework; things are never 'valuable' \textit{simpliciter}. This being the case, we should be skeptical about universalistic statements of value. This skepticism renders suspect any theory that makes strong claims about certain kinds of relations among agents (for example) being valuable.

However, it may be useful to ask what a \textit{minimal} addition to this value theory would look like, and how such an addition (\textit{qua} weak premise, with which most of us would agree) would change the conclusions that we can draw from our convention-based account of moral constraint. What we propose adding here to our conventionalist account, in order to thicken our current theory of value, is roughly an unsupported (though

\textsuperscript{46} See (Rawls 1971, p.19).
\textsuperscript{47} I owe the idea of adding a "thin" value theory to Hardin (1988), though he makes very different use of the concept.
hopefully plausible) moral premise. That is, we can only expand our theory of value by stipulating that something (other than the objects of individual preference) is valuable. In so stipulating, we would in effect be claiming that the object of such value would be valuable to – i.e., ought to be sought by – all persons.

The literature of moral and political philosophy is full of examples of values that we might add to our conventionalist account. We might, for example, thicken our value theory by stipulating that equality of opportunity is valuable. Indeed, in North America this claim would be almost universally accepted. The political Left and Right would surely disagree over just what equality of opportunity required (formal access to all jobs and political offices? equal access to education? affirmative action programmes?), but what little agreement might be found could be enough. The admission of this one non-individualistic, non-instrumental value could serve as an additional filter, a criterion by which to judge conventions. One or more such minimally controversial selection criteria might serve to rule out a range of sub-optimal or pernicious conventions, thereby bringing the results of our conventionalist approach more into line with existing moral intuitions.

Alternatively, rather than adding a value premise, we might add a procedural constraint to our account of morals. For example, we might accept the need for a Rawlsian veil of ignorance, or stipulate – with Scanlon (1982, p.110) – that systems of moral rules must be the based upon “informed, unforced general agreement.” Such a procedural constraint might at some deeper level be based upon a value claim of some sort (e.g., Kantian claims about the inherent value of all rational agents, which inspires Rawls’s veil). Or it might rather be a claim about what it is that defines or is constitutive of morality or
justice or right action. Either way, we would then have a non-conventional standard by which to judge the output of our convention-based account.

Notice an important difference between this proposal and Rawls’s assumption about equal moral worth. I propose that we should — unlike Rawls — examine first how far a convention-based account can take us, and then add in moralized premises afterward. This seems cleaner methodologically, and has the advantage of not building moralized premises in too soon. Since such moralized premises will necessarily weaken our conclusions, they ought not to be folded into the theory too soon, lest they render the whole theory suspect (as some would argue Rawls’s assumption of equal moral worth does).

I conclude, then, that the idea of a normative theory that accords a significant — perhaps even dominant — role to conventions warrants further investigation. Such an approach, with its focus on agents’ interest in coordination, provides a plausible account of moral motivation. Conventions carve out a reasonable range of constraints on self-interested behaviour. Given further refinements in our understanding of how conventions work, a convention-based approach to normative ethics holds considerable promise.
Chapter 6:

Beyond Symmetry
By now it will be obvious that a convention-based moral theory will not readily produce a set of obligations congruent with those embodied in our everyday morality. In this Chapter, we will focus on the ability of conventions to countenance asymmetries. Some of these asymmetries would likely be called, from the point of view of everyday morality, injustices. Later, in Chapter 7, we will deal more generally with inconsistencies between convention and everyday morality and raise a more general defence against this challenge.

Not all conventions will be egalitarian in nature; only in an important but limited range of examples are conventions fully symmetrical. Thus conventions as patterns of behaviour can embody serious inequities. And if we are right in holding that conventions generate obligations, then conventions can imply obligations to accede to injustices. While a convention-based approach can, of course, support an egalitarian norm (i.e., where appropriate conventions exist), this approach cannot be expected reliably to generate such norms, nor automatically to repudiate inegalitarian norms or situations. Given the strength of the egalitarian bent in modern moral theory, and the pervasiveness of egalitarian norms in (at least) Western culture, this conflict needs to be addressed. In this chapter, I examine two different kinds of asymmetry that can be supported by conventions. I argue that this ability of the conventionalist approach to integrate asymmetrical relationships may be a virtue, rather than an unmitigated failing. I suggest that the egalitarian impulse in modern moral theory is a bias, and that the notion of equality is not a necessary ingredient of a moral system. Further, building equality in as a premise unnecessarily limits the range of systems that can be called 'moral.'
Two Kinds of Asymmetry

Two different sorts of asymmetry must be dealt with. The first sort involves the inherent asymmetry of some strategic situations, i.e., situations in which one agent has more to gain from interaction than the other. In some such asymmetrical situations conventions will be possible, but will replicate inequalities in ways that seem problematic. For example, if two business partners have an unequal stake in cooperation, a convention could arise that involved unequal contribution to the joint effort. In other such asymmetrical situations, the asymmetry will be so great that one agent can gain nothing from the other agent's constraint or contribution, and thus has no interest at all in cooperating. In such situations, conventions will not be possible, which again raises questions about the adequacy of a convention-based approach.

The second sort of asymmetry involves asymmetrical payoffs in a convention that governs a symmetrical strategic situation. That is, it involves situations in which agents come to the interaction with an equal stake in the cooperative venture, but end up benefiting to different degrees. We can identify two forms of such situations. In one such form, agents have an equal stake in cooperation, but the available outcomes are such that there will always be some agent who benefits more than the other(s). This will be the case, for example, in any joint venture that requires that some person or another take a well-paid leadership role. Thus whatever convention arises will involve asymmetry, since not all agents will benefit equally (even if each of them had, in principle, an equal chance of being the one who benefits most).

The second form of this general type of asymmetrical convention is to be found in
cases in which the game being played is symmetrical, but some asymmetrical pattern of
behaviour has (for whatever reason) become a regularity. Such asymmetrical conventions
can be stable, despite their historical arbitrariness.

Asymmetrical Situations

First, we will examine cases in which the strategic situation itself is asymmetrical.
So far in previous Chapters we have dealt with conventions that mediate interaction among
agents playing similar roles. That is, we have dealt with what we might call symmetrical
interactions. For example, two nations engaged in a limited war are, per hypothesis, in
exactly the same situation. They have the same interest in the continuation of the
convention that puts limits on the ways in which they carry out their conflict. And we all
stand to benefit to the same degree from the bank-queue convention.\(^1\) Similarly, the
members of a society with a truth-telling convention all have (most of the time) a similar
interest in being told the truth, and a similar interest in being able to deviate unilaterally.
Technically, this is reflected in the fact that the parties involved have symmetrical
preference orderings over outcomes. It is this sameness that produces the shared interest in
coordination that sustains the convention.

So far we have said very little about the role of conventions in the wide range of
cases in which interaction is asymmetrical. Strategic interaction that is asymmetrical across
agents, as reflected in the agents having asymmetrical preference orderings over outcomes,\(^2\)

---

\(^1\) Even though, as noted in Chapter 3, the benefits will be unequal on any given day: whoever is at
the front of the line benefits most.

\(^2\) That is, Player 1’s ranking of the outcomes is not just the mirror image of Player 2’s ranking. Note
that such games are typically ignored by traditional Game Theory, which assumes that the agents
modelled are similar in all significant respects.
will often be the result of a power imbalance (a difference in natural or social endowments) between agents. In some of these situations, the payoffs available will be such that despite the asymmetry, all of the agents involved still have an interest in coordination; in these situations, conventions may still arise despite the asymmetry.

For example, a Fortune 500 firm may have something to gain (e.g., lowered transaction costs) from habitually purchasing a particular product from a particular small supplier. Thus it may be worth some effort for the larger company to maintain an amicable relationship; the larger company may therefore choose to ignore small errors in shipments, or trivial breaches of the wording of the sales contract, and so on. To the smaller company, on the other hand, the continuing relationship may be utterly crucial; to them, smoothing out problems with such an important customer could mean the difference between survival and obsolescence.\(^3\)

<table>
<thead>
<tr>
<th></th>
<th>Microsoft</th>
<th>Don't trade</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mom &amp; Pop Firm</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trade</td>
<td>2</td>
<td>n.a.</td>
</tr>
<tr>
<td>Don't trade</td>
<td>n.a.</td>
<td>-4</td>
</tr>
</tbody>
</table>

Figure 6.1: “Asymmetrical Trade Relationship”
Payoffs here are cardinal rankings.\(^4\) Thus while trading is better for Microsoft than not trading, for the Mom & Pop Firm trading is much better than not trading. (Since it takes two to trade, the payoff for unilateral trade is undefined.)

\(^3\) In technical terms, this difference may be expressed as an asymmetry in the two companies’ cardinal rankings of various outcomes.

\(^4\) Ordinal rankings only rank outcomes in the order in which they are preferred. Cardinal rankings indicate more: they indicate degree of preference.
As a result, the small supplier may be willing to ignore rudeness on the part of the purchasing department at the larger company, last-minute changes to orders, late payments, onerous penalties for late shipment, etc.\(^5\)

Thus a rather one-sided convention might arise, one that is beneficial to both (i.e., better than not doing business with each other at all) but which looks grossly unfair from the point of view of everyday morality. (This convention is represented by the \{1,3\} outcome in Figure 6.2.) The convention that arises will thus reflect, rather than mitigate, the power imbalance between the parties.\(^6\)

Another example of an unequally-beneficial convention would be a convention that arises between the owners and the employees of a third-world sweatshop.\(^7\) The workers in a

\(^5\) Dawkins (1989, p.250) calls asymmetries in the cost of failure the “life/dinner principle,” after one of Aesop’s fables according to which “the rabbit runs faster than the fox, because the rabbit is running for his life while the fox is only running for his dinner.”

\(^6\) Gauthier (1986, p.97) notes that the ideal conception of the market rules out all forms of interaction in which one party “does not benefit from interaction to the extent of his contribution.”

\(^7\) As Gauthier (1986, p.110 ff.) points out, Marxists will see this sort of exploitation as inherent in the operation of market economies. Gauthier notes that under perfect market conditions, the economic rent which capitalists are able to earn on labour will vanish as the supply of labour is
Haitian garment factory may accept very poor working conditions - a fourteen hour day, poor ventilation, etc. - because such employment is better than the alternative. Let us imagine such a garment factory, one in which workers are treated badly (by North American standards) but are not beaten or threatened or the like. We can imagine the following convention arising between labour and management: workers constrain themselves by not complaining to members of visiting international human rights organizations; in return, management constrains itself by abstaining from the threats and beatings to which managers of other factories regularly subject their employees. This lopsided convention is beneficial to both (i.e., better for the workers than unemployment would be, and better for management than having human rights workers pestering them would be) but which looks grossly unfair from the point of view of everyday morality.

Further examples of such asymmetrical situations might include the situation faced by a physician and her patient, or by a graduate student and her supervisor. In both cases, asymmetries of knowledge and power mean that one party has more to lose, and thus may be willing to accept higher costs in order to reap fewer or lesser benefits from cooperation. Therefore the conventions that arise to govern the relationship are unlikely to be egalitarian.  

In other asymmetrical situations, the asymmetries will be so severe that not all agents will have an interest in coordination at all. Such situations, unfortunately, cannot be brought into equilibrium with the demand for it. Gauthier also points out, however, that assumptions about perfect markets do not imply that exploitative relationships of this sort cannot arise in the real world.

---

8 Even more extreme examples would include the relationship between parent and child, the physically challenged and the able-bodied, and (perhaps an offensive comparison) a dog and its master. (Some might see the last example as a reductio; I do not, but will not defend that position here.)
remedied by conventions, which by definition must be sustained in part by a shared interest in coordination. Hume concurs: in his *Enquiry* (1975/1751, p.190-1; qtd. in Gauthier 1986, p.308), he argues that if we had amongst us a class of rational beings so feeble as to be incapable of affecting us, “the restraints of justice and property, being totally *useless*, would never have place in so unequal a confederacy” (emphasis original).

A situation calling for benevolence on the part of a wealthy individual, A, toward a very poor individual, B, is asymmetrical in this regard.\(^9\) If the two possible outcomes are one in which A gives $20 to B, and one in which A gives no money to B, the two agents will have symmetrical *ordinal* rankings of the outcomes (A’s most preferred choice is B’s least preferred choice, and vice versa), but will likely have very different *cardinal* rankings of the outcomes. This reflects the much greater stake that B has in the situation. Giving away $20 is almost negligible for A; receiving $20 would be an enormous boon to B. In this case, A has no interest at all in coordinating with B (since B has no choice to make, or at least no choice that materially affects A). Such being the case, this kind of situations will never be subject to convention. A moral theory that derives obligation from convention, therefore, will not be able to produce an obligation for A to give money to B, even when the money at stake is negligible to A and will benefit B greatly.\(^10\)

Of course, it might be thought that there *is* a way in which a convention-based approach could generate such obligations of beneficence. After all, it could be argued that a convention of mutual aid *would* be mutually beneficial in that it would serve as a kind of

---

\(^9\) Such a situation might not strictly be called ‘strategic,’ in that one of the agents may not have any real choice to make. Here the tools of game theory still seem useful, even though the situation does not strictly speaking have the structure of a game.

\(^10\) Ignatieff (1998, p. 20) refers to the charity of strangers as “that most fearful dependency relation.”
insurance scheme. This would amount to a Rawlsian argument; it seems plausible that, from behind a veil of ignorance, rational agents would choose such an insurance scheme, particularly if it involved having to give aid only when the inconvenience is small and the benefit great. However, such appeals to hypothetical agreement are too optimistic. It is hard to see what interest real, affluent people would have in an ethic (or even a rationality) that forced them to constrain themselves in ways that are not in their interests.

Some will see the example given above as constituting damning evidence against a convention-based approach. How can it be morally permissible not to help those in need, when it costs so little? Yet others will be unfazed: many regard giving to charity as supererogatory. Few of us make more than token donations, for example, to charitable causes. Most of us pass by panhandlers without pausing to help. So many will not find much to worry about in this example. But if we are to do a good job of defending a conventionalist account, we must examine its most seriously counter-intuitive implications.\(^\text{11}\) Perhaps the example cited above was insufficiently challenging – i.e., insufficiently counter-intuitive – because it involved particular individuals transferring small sums of money, and because we left unspecified the reason why the poor man was poor and the reason why the rich man was rich (perhaps he got rich by exploiting his now-poor neighbour). Or more likely it was insufficiently challenging because it involved a

\(^{11}\) Harman (1983, p.124) points to the fact that we generally think that the duty not to harm is greater than the duty to help as evidence for the conventional nature of justice. He suggests that this difference “is a consequence of the fact that morality represents a compromise between people of different powers and resources, since everyone benefits from a strict duty not to harm others, but only the poor and weak would benefit from an equally strict duty to help others avoid harm.”
disparity in 'mere' material wealth, rather than a disparity in basic liberties.\textsuperscript{12} A more challenging example might involve an opportunity to redistribute rights and liberties rather than mere money. Could a convention-based account of moral obligation generate duties to move toward a more egalitarian distribution of rights?

Let us take as a more challenging example a situation in which the asymmetry involved some people having a greater range of \textit{rights} than others. Imagine, for example, a feudal society consisting of peasants (with only minimal legal rights) and aristocrats (with comprehensive rights, both legal and political). Imagine also that the peasants do in fact benefit from the existing social arrangements; perhaps they even think of their lives as being quite good lives. Could members of the aristocracy have an obligation to improve the lot of peasants in general, from a conventionalist perspective? Or could particular aristocrats have an obligation to improve the lot of particular peasants, say by means of transferring to them some rights or resources? If, \textit{per hypothesis} the aristocrats have nothing to gain from improving the lot of the peasants, then the logic applied in the business example above seems to apply equally well here. The members of the advantaged class cannot have a strictly conventional obligation to give up some of their rights to members of the disadvantaged class. The \textit{status quo} is mutually (though not equally) beneficial; the proposed change, again, is not.\textsuperscript{13}

Next, let us move to a harder example still. Imagine a situation in which again the asymmetry involved has to do with people's range of effective rights, but this time imagine

\textsuperscript{12} The greater importance of basic liberties than of social and economic status is reflected, for example, in Rawls's (1971) assignment of lexical priority to his 'principle of equal liberty' over his 'difference principle.'

\textsuperscript{13} I have not attempted to provide, here, a theory of bargaining. It seems clear, though, that a convention-based bargaining theory would use the \textit{status quo} as its baseline.
a society of slaves and masters. As with the previous example, we will assume that the slaves do benefit overall from their relationship with their masters. They are not mistreated – perhaps some are even fond of their masters – but they are slaves none the less. Could members of the master class have an obligation to improve the lot of slaves in general, from a conventionalist perspective? Or could particular masters have an obligation to improve the lot of particular slaves, say by means of transferring some rights or resources? If, per hypothesis the masters have nothing to gain from improving the lot of the slaves, then the logic applied in the charity example above seems to apply equally well here. Just as the rich man was unmotivated to give the poor man $20, the slave-owner will find little reason to transfer rights to her slaves. Without there being any advantage to doing so, it seems again that the members of the advantaged class cannot have a strictly conventional obligation to give up some of their advantages to members of the disadvantaged class. The status quo is mutually (though again not equally) beneficial; the proposed change is not.

Note that this does not constitute a moral argument in favour of slavery, or even in favour of maintaining the status quo. Rather it is an argument about what sorts of moves it will be possible to justify in particular situations (keeping in mind that, as argued in Chapter 5 above, actions are never justified simpliciter – they are always justified to someone.)

We will move now to consider a different type of asymmetrical convention – namely asymmetrical conventions between individuals who do have the same interest in cooperating – before attempting to deal with the challenge posed by both of these kinds of

---

14 This example – and its use as a difficult challenge to a moral theory based on rational choice – is borrowed from Gauthier (1986, Chapter VII).
Asymmetrical Conventions in Symmetrical Situations

Some conventions are themselves asymmetrical in that they produce different payoffs for different agents, but are better for all concerned than other available outcomes. That is, agents might come to the interaction with an equal stake in the cooperative venture, but the available outcomes may be such that there will always be some agent who benefits more than the other(s). This will be true, for example, of any convention that 'solves' a game of the Battle of the Sexes. In such a game, both parties prefer acting according to convention over failing to coordinate, but the parties differ over which of the two (or more) available coordination points is the best. For example, in order to communicate with each other, a (mostly) bilingual Anglophone and a (mostly) bilingual Francophone need to decide whether to converse primarily in English or in French. Either convention would be Pareto superior to trying to carry on the conversation in both languages at once (though the two might enjoy the challenge for a short while). But (let us assume) each of these people would rather converse in her mother tongue.

![Figure 6.3: "The Battle of the Linguistic Groups"

<table>
<thead>
<tr>
<th></th>
<th>Speak French</th>
<th>Speak English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Franco-phones</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speak French</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Speak English</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 6.3: "The Battle of the Linguistic Groups"
Thus the Anglophone would rather that both follow one particular convention – i.e., speak English – while the Francophone would prefer the convention of speaking French.\footnote{Another example would be a situation in which two manufacturers want to make their computer components compatible, but disagree over whose proprietary technology to use to accomplish the feat.} Note that unlike the situations discussed in the previous section, this situation \textit{itself} is not asymmetrical. Each participant has an equal stake in agreement. That is, the payoff matrix is symmetrical, but both of the Pareto optimal outcomes are asymmetrical as far as payoffs to the two agents are concerned. Indeed, in the case outlined here, the choice between the two Pareto optimal outcomes is a constant-sum game. No matter which Pareto optimal outcome becomes the convention, there will be a winner and a loser.

Where conventions are asymmetrical in the sense exemplified by the Battle of the Sexes (i.e., different payoffs to different agents, but better than non-agreement) the asymmetry is, at first glance, unavoidable. In simple cases, at least, all available Pareto-superior outcomes will be asymmetrical. Thus if our convention is to speak English, and if you prefer French, both of us switching to French would be better for you, but would not in general be any more fair. As is often the case, iteration would help: if we converse on a number of occasions, a more strictly egalitarian convention of switching between English and French on alternate occasions becomes available.\footnote{Note that this solution of alternating is less plausible for questions of technical standards. Using Mac protocols on odd days and Microsoft protocols on even days would presumably be enormously inefficient, to the detriment of all concerned.} For such simple cases, we might point out that even though the outcome will necessarily be better for one party, we can at least achieve procedural justice by ensuring that the process by which the convention is chosen is impartial. For example, we could follow Gauthier’s (1986, p.120) suggestion, and
toss a coin. The payoffs to each of us under either convention (alternating and coin-tossing) are the same: each of us is satisfied (or dissatisfied) half the time. But in considering attempts to find alternatives that seem more fair, we should recall that this is a constant-sum game – any gain for you is a loss to me. I will thus be prima facie unmotivated to switch from the English-only convention once it has become established. You might of course try to change the convention by changing your own behaviour, but your unilateral deviation from the English-only convention hurts us both. Whether the bad non-cooperative payoff motivates me to accede to your change in convention or instead motivates you to return to the English-only convention will depend on which of us has a greater stake in the interaction.\(^{17}\)

A final form of asymmetrical convention is to be found in cases in which the game being played is symmetrical, but some asymmetrical pattern of behaviour has (for whatever reason) become a regularity. Imagine, for example, the following idealized situation, suggested by Sugden (1990).\(^{18}\) Two agents – a man and a woman – must decide how to divide some good. If they fail to agree, the good goes to waste. Let us further imagine that it becomes usual (say, for some arbitrary historical reason) for the woman to demand 2/3 of the good, and for the man to demand just 1/3 of the good. If this tradition is commonly known, then it will be rational for women to continue to demand 2/3 and men to continue to demand just 1/3. That is, if a man knows that women typically claim 2/3, then it is rational for him to claim 1/3 whenever he interacts with a woman; and if a woman knows that men typically claim just 1/3, then it is rational for her to claim 2/3 when she interacts with a

\(^{17}\) That is, it will depend on our ordinal rankings of the outcomes.

\(^{18}\) Sugden (1990, p.770) notes that this game is a version of the bargaining game used by Gauthier, which is in turn borrowed from Nash (1950).
man. This pattern of behaviour constitutes an asymmetrical convention. Although women benefit more from it than do men, all are better off than they would be in the absence of any convention (since without a convention they would often fail to agree, and so the good in question would often go to waste). Asymmetries of this kind need not involve an instantaneous division of some good. They can instead involve one party to an interaction failing to contribute, in some patterned way. For example, agent A may put up with agent B defecting at every (say) fourth interaction, if there is a threat that intolerance on the part of A will result in B shifting to more frequent defection. 19

As (Sugden 1990, p.781) has noted, some people have argued that “if a convention systematically favors one group of people at the expense of another, some individuals in the disfavored group will refuse to go along with the convention in an effort to change it.” This may of course happen, but as Sugden (p.781) rightly points out, “what is at issue here is whether it would be rational to behave in this way, when ‘rationality’ is defined in terms of the efficient pursuit of individual interests. The whole point of contractarian theory is to derive moral principles from such an instrumental conception of rationality.” Our project is similar to the contractarian project in this regard. A convention embodying unequal benefit will nonetheless be beneficial to all; moving away from such an asymmetrical convention will not be beneficial to all, and thus will be undermotivated.

Thus there may be conventions that embody various kinds of asymmetry. At least some of these asymmetries will look like injustices, which is worrisome given our contention that conventions have moral force. The result (for at least some of the examples

19 In game-theoretic terms, this point is best represented by the infinite number of asymmetrical equilibria of the Iterated Prisoner’s Dilemma.
cited above) is a disconcerting moral endorsement of what is *prima facie* an immoral situation. Further, conventions will not generate obligations to help those from whom we have nothing to gain. The result here is a disconcerting moral silence with regards to a wide range of tragic situations: a conventionalist morality will not, for example, prompt North Americans to send aid to Somalia or Rwanda. That is, conventionalism fails to remedy certain ‘unjust’ situations.

**The Conventionalist Response**

Our response to these charges will be twofold. Our first response is that a convention-based account does have the resources at least to lessen the impact of these potential asymmetries. Our second response will be to argue that the ability to countenance certain asymmetries should actually be considered a virtue in a moral theory.

We will be begin with ways in which at least some asymmetries can be mitigated under a conventionalist approach. First, it should be pointed out that the apparent injustices of some of the situations described above would be moderated by the existence of more general conventions. For example, to the extent to which there is a general societal *convention* of egalitarianism, then that convention could reasonably be expected to act as a constraint on the kinds of conventions that can arise in particular interactions between individuals, even when those two individuals differ in terms of their natural or social endowments. I return to the possibility of such general, over-arching conventions in Chapter 7.

Further, it should be noted that even though conventions between the rich and the poor cannot generate obligations of beneficence, conventions among the rich might be able
to. That is, if the affluent segment of a society decides (for whatever reasons) that it ought (rationally or morally) to render aid to the indigent segment of society, then a convention may arise that implies that particular affluent individuals do have duties of beneficence.\(^{20}\) The simplest example would be the plausible case in which the affluent members of a society feel that it is in their collective interest to ensure a basic minimum of welfare to the least advantaged members of their society. It might be felt, for example, that ensuring that each member of society has the basic necessities of life is the best way to reduce crime; or to ensure a healthy and at least minimally skilled working class.\(^{21}\) Given such a shared belief, the affluent segment of society might try to achieve their goal by one of two methods. First, they might empower their government to collect taxes in order to provide a social safety net. Alternatively, they might rely on individual charity, and evolve a convention that implied obligations to help those less fortunate. Were the latter route to be taken, then affluent individuals would have an obligation to help the indigent. This would be a true convention: each would have an interest in achieving their goal of (instrumentally) aiding the poor, but no individual would feel obligated to help (and indeed, her charity would do little to achieve this shared goal) if her fellows did not do so as well. Note that such a conventional obligation would be an obligation to other members of the affluent segment of society – to other participants in this collective action – but the object of the obligation would be those less fortunate members of society who stood to be helped. Thus

\(^{20}\) Hardin (1988, p.56) suggests that if group A acts beneficently toward group B, each member of group A has two kinds of obligations, namely obligations to members of group B and obligations to other members of group A. I propose a similar basis for physicians’ professional obligations later, in Chapter 8.

\(^{21}\) Alternatively, the affluent might feel that they need to improve the situation of those less well-off in order to quell their own natural, but apparently irrational, sympathies.
conventions can at least sometimes help those who are not party to them.

As a second type of response to the charge that conventionalism countenances problematic asymmetries, I suggest that this ability of the conventionalist approach to integrate asymmetrical relationships may be a qualified virtue, rather than an unmitigated failing. We can reach this more optimistic conclusion by pointing out that the egalitarian impulse in modern moral theory is a bias; that is, from a functionalist point of view, egalitarianism is a substantive moral thesis, rather than a necessary ingredient of a moral system. If Gauthier (1986), for example, is right, then impartiality is in fact an important determinant of the stability of a moral order. But the details of this story are contingent matters: we cannot establish that all moral codes must be impartial simply by stipulation. Building equality in as a premise unnecessarily limits the range of systems that can be called ‘moral.’ Indeed, history (and current events both at home and abroad) clearly shows that most people think that we can be moral without being impartial. I am of course not making a moral argument to the effect that we should countenance bias or discrimination: as Western egalitarian liberals, we should abhor such things. Yet if we accept the premise that morality is an artifice, a piece of social technology used to make group living more congenial, then it becomes largely an empirical question whether an egalitarian system is best suited to meeting the needs of social existence in particular times and places. We ought not to beg the question in this regard. Social situations embodying asymmetries can still have the characteristics needed to attribute moral force to the conventions governing them. Let us begin with a relatively innocuous example, and move from there to more challenging ones.

Barry (1989, p.291), for example, refers to impartiality as the very “core of morality.”
We will begin by returning to the case of *symmetrical* situations resolved by asymmetrical conventions. The example cited above involved an Anglophone and a Francophone needing a convention to determine the language in which they would converse. It was noted that no matter which of the two (pure) conventions came into effect, one or another of the two linguistic groups would be disadvantaged. Let us assume that a particular standard is in place: “speak French.” And in order to move one step further from toy problems, let us assume that rather than just two individuals, an entire community – half Anglophone, half Francophone – is subject to this standard. *Per hypothesis*, all members of the community benefit from there being a convention. Let us assume, for the sake of argument, that sorting out which language to speak on a case-by-case basis (i.e., whenever two individuals meet) raises transaction costs intolerably. Let us assume further that although they all agree that the convention is beneficial, the Anglophones in the community would still in principle prefer that everybody speak English. If, as would likely be the case, the existing standard constitutes a regularity of behaviour; and if that regularity can be attributed to the interest which all parties concerned have in coordinating, and to the expectation on the part of any one individual that others will also conform; then this regularity – this standard – will constitute a convention in the sense we are interested in.

Now while we might regret the fact that no more egalitarian arrangement seems possible, we would surely agree that in such a situation the standard, “Speak French” would be taken by the community as a whole to have moral force. Clearly, to deny moral force to this convention on the simple grounds that it benefits some more than others would be overly harsh; to insist on absolute equality of condition in such a case would be to deny this community access to either Pareto-optimal solution. In endorsing – giving moral weight to
– this form of asymmetry, conventionalism seems quite reasonable: surely our moral intuition in this matter must be that it embodies no real injustice. If indeed “ought” implies “can,” then the absence of an accessible egalitarian alternative prevents us from calling this convention unjust. Conventionalism might, however, still be criticized for stifling innovation. If the status quo ("Speak French") has moral force, then there seems to be no reason, and indeed no justification, for seeking new solutions through, for example, the development of new technologies.²³

Let us move now to a more troubling set of examples, and see whether the ability to integrate asymmetrical relationships can still be considered a virtue. As noted above, a convention-based moral theory seems to replicate, rather than mitigate, pre-existing inequalities. For example, it was noted above that if two agents have an unequal interest in cooperation, they can arrive at a convention that is unequally beneficial: the party with proportionally less to gain from cooperation may be able to insist upon a greater share of the cooperative product. Indeed, an unequal division of the fruits of cooperation may constitute a stable convention even if it is the product of some historical accident, rather than the result of some initial difference in bargaining power. While we may express concern at this pattern, it would be inappropriate to deny that such conventions should be called “moral.” After all, they do serve to improve social interaction. Indeed, we could run a reductio on the claim that any division other than an egalitarian one is either immoral or amoral. Just how scrupulously egalitarian does an arrangement have to be before it can

²³An example of a technological solution to this problem would be the development of a new language. The costs of developing a new language, of course, might be such as to make this solution Pareto-inferior to the existing convention. Note the abject failure of Esperanto to gain currency as a diplomatic lingua franca.
have moral force? Surely there are situations in which any number of bad arrangements should be accorded moral force, on the grounds that they prevent slipping into truly terrible situations. An analogy: a benevolent dictator may be tolerated because the population can imagine much worse.

From the perspective of a set of individuals trying to move out of a Hobbesian state of nature, it seems foolish to insist upon equal division if some individuals are so situated that they can realistically demand more. Take leadership hierarchies as an example. From an evolutionary point of view, leadership hierarchies are a crucial innovation. It is a truism of the social sciences that individuals can produce more by working together than they can by working separately; simply put, cooperation is efficient. But undirected joint effort requires much of individual agents. Egalitarian teamwork requires that each agent have the cognitive capacity to recognize and prioritize goals, to anticipate other agents’ actions, and to modify her own actions as her perception of the environment changes.24 Even where cognitive capacity is relatively plentiful, undirected individuals may come up against the logic of collective action (as outlined in Chapter 4 above). Even chimps and wolves, on the other hand, have ‘learned’ that centrally coordinated efforts result in successful conduct of warfare and hunting of prey. Further, many technologies — from 18th Century sailing battleships to 20th Century multinational corporations — require complicated command hierarchies. Where such efficiencies require inequality, it seems foolish to be steadfastly egalitarian. Even so prominent an egalitarian as Rawls (1971) admits that certain

24 Social insects may seem to be a counter-example: they are hugely successful at coordinating the activities of tens of thousands of individuals despite their lack of true leadership. However, this success can be attributed to the fact that individual ants, for example, are genetically so similar to each other that they function more like the cells of an individual organism. They thus don’t have divergent interests to coordinate.
inequalities can be tolerated where these are efficient.  

There is, of course, a significant literature concerning whether justice (or morality more generally) is a question of mutual advantage or rather a question of impartiality. Those who, like Rawls, focus on impartiality usually begin with an assumption about the equal moral worth or inherent moral status of all agents. According to Kymlicka (1990, p.132), “[m]ost political philosophy in the Western tradition...shares the...view that there are obligation-generating rights and wrongs which all persons have a duty to respect.” A convention-based account runs afoul not only of common-sense, but also of any theory that accepts that there are natural duties or moral values in the world. Kymlicka (1990, p.131), for example, charges that “the mutual advantage approach is not an alternative account of justice, but rather an alternative to justice.”

This categorical insistence upon the impartial nature of justice is odd, given that one meaning commonly attributed to the word ‘justice’ is something like the vague formal criterion of ‘each getting what he or she is due.’ Certainly that understanding of justice is consistent with a wide range of substantive principles. Kymlicka (1990, p.132) admits that the impartiality constraint is an assumption, supported only by something like intuition. And Barry (1990, p.282 ff.), despite trying, fails to find more convincing support for impartiality. Rather he is forced to point to the need which humans apparently have, and feel, to justify their actions, and to do so by reference to reasons that are appealing to their fellows. Barry argues half-heartedly that this should be counted a criterion of rationality; but it seems clear that precisely what sorts of reasons will succeed in justifying one’s actions to one’s fellows should be seen as a contingent matter, rather than something that

---

25 I refer here, of course, to Rawls (1971)'s famous “difference principle.”
should be made a matter of definition. If an agent in a given situation is only able to succeed in justifying her actions to her fellows to the extent to which those actions are impartial, then so be it. Given an egalitarian convention, she will need to reason impartially in deciding how to act; if she fails to reason impartially in the face of such a convention, she will surely be unable to justify her actions to her fellows after the fact. But why assume that an impartial standard is the one that will apply? My point here is not, of course, to provide an argument in favour of an inegalitarian standard of justice; nor is it to give a justification of inequities past, present, or future. It is simply a claim that the particular standard of justice governing particular issues in particular settings should be seen as a matter of convention.

Conclusion

In this Chapter I have discussed ‘asymmetrical’ conventions, conventions in which different parties benefit to different degrees. It seems that a conventionalist approach to ethics must give prima facie endorsement to at least some such asymmetries. I have argued that while a convention-based approach cannot forbid such asymmetries, such an approach at least presents the possibility of their being higher-order conventions that would mitigate their effects. I have also argued that to call such asymmetries – such inequalities – injustices is to accept a certain conception (common though it may be) of what justice requires. Justice qua equality (either equality of consideration or equality of situation) is but one substantive theory of justice. It ought not to be taken as constitutive of ethics itself.
Chapter 7:

The Limited Aspirations of a Convention-Based Account
In this chapter, I defend against the charge that a convention-based approach fails to match adequately the tenets of everyday morality. To do this, I offer two lines of defence, each of which involves limiting aspirations. The first line of defence involves modesty about the role of arguments from convention in an overall ethical theory, and argues for a place for conventions in the range of arguments used by applied ethics. The second line of defence is more optimistic about the role of conventions, but proposes more modest goals for moral theory itself.

In Chapter 5, I discussed two potential criticisms of a convention-based account. The first was a concern about conservatism: I argued that a convention-based account is not nearly as conservative as, say, a norm-based account would be. Further, I pointed out that while conservatism may at times appear reactionary, it has at least the virtue of being cautious. Conventions are, after all, mutually advantageous; a bad convention will be better than no convention at all. While this is no reason not to question the sufficiency of any given convention, in pointing out that conventions constitute a useful class of constraints we warn against unthinking change.

The second potential criticism had to do with the possibility of pernicious conventions. In this regard, I argued that while some conventions may well be detrimental to ‘outsiders,’ there is nothing incoherent about the idea of in-group moralities. Yet even if these defences are taken as being successful, we are still faced with the question of whether a theory of ethics that bases obligation on convention conforms closely enough to the tenets of everyday morality. In Chapter 6, we dealt at length with one particular aspect of this
challenge, namely conventionalism’s apparent ability to accommodate – and even to generate – situations that, from the point of view of our common moral intuitions, seem unfair. In this Chapter we shall discuss further differences between conventionalist morality and everyday morality, and attempt to defend a convention-based account against this troubling criticism. It should be noted that the divergence between the standards that can be generated by convention and the standards of everyday morality illustrates the problem with using the terms ‘conventional morality’ and ‘everyday morality’ as synonyms.

It is generally seen as one measure of the worth of a moral theory that it goes some distance toward, as Rawls (1971, p.579-80) puts it, matching “the fixed points of our considered convictions better than other familiar doctrines.” In this section, I admit that a convention-based theory appears to fall short in this regard. I then argue that despite this apparent weakness, a convention-based account has a legitimate role to play. It is tempting to reply to this criticism by asking whether the existing mainstream moral theories fare any better in this regard. Utilitarianism, for example, is exceedingly popular, despite the famously counter-intuitive conclusions to which it can be driven.¹ Kantian absolutism is similarly suspect. Instead, I propose two less schoolyard-like lines of defence. First, I suggest that even if conventionalism fails to provide a convincing all-purpose moral theory, arguments from convention may still deserve a spot among the short list of considerations utilized by a pluralistic approach to applied ethics. Second, I point out the role of moral sentiments in filling at least part of the gap left between common-sense morality and the obligations that a convention-based approach can sustain. Finally, I argue that it may be

¹ Perhaps the most famous critique of Utilitarianism along these lines is Parfit’s (1984, p.388) “Repugnant Conclusion” argument. The best defence against such charges is likely that found in (Hardin, 1988).
possible to see a convention-based theory as fully credible on its own, given a more realistically modest estimation of what moral theory can actually accomplish.

**Conventions and Everyday Morality**

It will be clear by now that a convention-based moral theory faces certain limitations with regard to its ability to generate obligations that correspond with the tenets of everyday morality. Conventions will not reliably generate, for example, all of the duties of beneficence that would be required if we want a moral theory to be brought into something approaching reflective equilibrium with common moral intuitions. Conventionalism, it seems, is subject to the same criticism that Locke levelled at Hobbesians. Gauthier (1986, p.268) quotes Locke as holding that a Hobbesian “…will not easily admit a great many plain duties of morality.”

Conventions cannot readily generate, for example, obligations to offer even minimal aid to strangers starving in foreign lands. On our model, obligations are taken as having their source in an interest in coordination. Problems of beneficence are constant sum games, offering no opportunity for coordination. Situations typically seen (by everyday morality) as calling for charity are seldom ones in which the potential benefactors have anything concrete to gain from helping; in most cases, it is unlikely that the tables will ever be turned such that present day beneficiaries have the opportunity to repay such kindness.

Nor, for example, can conventions easily generate more than minimal protections for non-human animals. Indeed, it is doubtful whether a convention-based morality could

---

2 Gauthier (1986, p. 268) notes that elements of his own theory “do not correspond to the ‘plain duties’ of conventional morality.” Note the troubling use of the term “conventional morality” to describe what I have called “everyday morality.”
include any limitations on our treatment of animals (subject to a few minor qualifications involving companion animals,3 and worries about the apparent connection between cruelty to animals and violence toward humans.4) Conventions almost certainly could not generate even the limited protections which farm animals in North America enjoy today.5 As Gauthier (1986, p.268) notes, “[a]nimals, the unborn, the congenitally handicapped and defective, fall beyond the pale of a morality tied to mutualism.” In fact, Harman (1983, p.124) points to the fact that we generally accord animals lesser (if any) moral status as evidence for the conventional nature of morality. He writes (p.124), “Our moral conventions favor people over animals because they are conventions arrived at by people for people.”

Further, a convention-based moral theory offers support for – i.e., it justifies – problematic asymmetrical moral relationships. Not all conventions will be egalitarian in nature. Indeed, as indicated in Chapter 6 above, only a limited (albeit important) range of examples are fully symmetrical. If mere conventions can generate obligations, then conventions as we have defined them can imply obligations to accede to great injustices. In Chapter 6, I also argued that this ability of the conventionalist approach to integrate

3 i.e., some animals might end up being the beneficiaries of a convention among humans. Thus property conventions will protect farm animals and companion animals from certain abuses (at the hands of people other than their owners). There might also be a weaker convention-based argument against me abusing my own animals. In particular, if my cruelty to my animals inflicts psychological harms on my neighbours, then my abstention from such cruelty might be part of a general convention involving abstention from a range of distasteful activities. Such activities might range from cruelty to animals to the public use of vulgar language. Thus we might derive a limited set of rights for animals from a (conventional) principle of respect for persons. For an evaluation of the limits of such arguments, see Carruthers (1992).

4 For example, Anderssen (1999, p.A21) cites American studies suggesting that “almost 40 percent [of animal abusers] had committed violent crimes against people” and that “30 percent of convicted child molesters admitted abusing animals when they were children or teenagers.”

5 See, however, Rollin (1993) who claims that the dominant ethic of traditional animal agriculture is one premised on mutual benefit.
asymmetrical relationships is a virtue. I argued that the egalitarian impulse in modern moral theory is a bias, and that the notion of equality is not a necessary ingredient of a moral system. I argued that building equality in as a premise unnecessarily limits the range of systems that can be called 'moral.' Whether or not my argument in Chapter 6 was successful, we are left with the fact that rough egalitarianism is part of everyday morality. And while a convention-based approach can, of course, in principle support such a norm (i.e., where appropriate conventions exist), such an approach cannot reliably be expected to generate such norms, nor to repudiate inegalitarian norms or condemn inegalitarian situations.

Finally, a convention-based account may not deal well with novelty. Since conventions are social regularities of behaviour, agents faced with situations for which there is no established pattern of conduct cannot look to convention for guidance. Situations that are novel either in the issues they encompass (e.g., situations involving the use of new technologies) or in the particular people they involve (e.g., when strangers from different cultures meet for the first time) cannot readily be seen as governed by convention-based obligations. Sugden (1989, p.93) suggests that conventions may spread from familiar situations to novel ones, so that we may find guidance for our behaviour in new situations by looking to the conventions that hold in analogous situations. But Sugden (p.94) further notes that the very concepts of similarity and analogy are subjective. All of us trying to choose the same analogy is itself a coordination problem, which prevents our obligations in novel situations from being clear.

Thus on an account which derives obligations entirely from conventions, agents in such situations cannot have obligations at all. Everyday morality, on the other hand,
attributes to norms the flexibility to deal with novel situations as extensions of familiar ones. I take it, for example, that everyday morality would by default grant certain rights to members of newly discovered civilizations. We hold anthropologists to roughly the same standards of research ethics with regard to isolated tribes as we do with regard to members of their own cultures and communities. And we would think it wrong of a Canadian business to set up a factory in an isolated Brazilian village and proceed to enslave the local populace. Similarly, everyday morality would draw on moral standards established to govern older modes of communication (e.g., face-to-face dialogue) to govern new modes of communication such as e-mail. In the latter case, we might close the gap between convention and everyday morality by arguing that the relevant convention is one of ‘civility of discourse,’ independent of the medium. But whether or not such an argument would succeed would depend, in part, on whether moral rules in general are thought to be most properly interpreted as applying narrowly or broadly. And whether moral rules are to be interpreted as applying narrowly or broadly is not to be determined on theoretical grounds. (There might be reason, however, to think that a standard of broad interpretation would be more conducive to social harmony. This is an empirical question.) Whether moral rules are to be interpreted broadly or narrowly is, I take it, a contingent fact – a matter of convention.\footnote{An analogy: I take it that the question of whether legal precedents are to be applied broadly or narrowly is itself a matter of precedent.} Thus there is at least the potential for there to be situations in which conventions offer little if any guidance.

Thus a convention-based moral theory endorses asymmetries, cannot readily generate obligations of beneficence, and deals poorly with novelty. How are we to deal
with these deviations from everyday morality? I sketch here two responses to this challenge.

1. The Toolbox Approach to Applied Ethics

My first defence of the apparent narrowness of a conventionalist approach is to argue that obligations based on convention constitute but one part of a pluralistic approach to ethics. Such a pluralistic approach is particularly in evidence in the various fields of applied ethics. Indeed, the approach which applied ethics generally takes to moral theory might best be described as eclectic – applied ethics literature most often seeks to apply (and elaborate) elements of several erstwhile conflicting moral theories. Writers in modern fields of application such as bioethics (Beauchamp and Childress 1994; Englehardt 1986), business ethics (Velasquez 1992; Poff and Waluchow 1991), professional ethics (Martin and Schinzinger, 1989; Yeo 1996), computer ethics (Johnson 1994) and environmental ethics (Piece and VanDeVeer 1995, Rodd 1990) typically rely on some combination of the mainstream moral theories. They tend to emphasize a range of considerations grounded in consequentialist theory, rights theory, theories of justice, and occasionally virtue theory or care theory. Velasquez (1998, p.128), for example, summarizes his introductory chapter on “Ethical Principles in Business” this way: “Our morality, therefore, contains four main kinds of basic moral considerations [utility, rights, justice, and care], each of which emphasizes certain morally important aspects of our behavior, but no one of which captures all the factors that must be taken into account in making moral judgments.”

---

7 The exceptions to this pattern are noteworthy. Singer (1993), for example, takes a wholeheartedly act-utilitarian approach to applied ethics. Narveson (1993) applies his libertarian framework to a range of moral problems.
Some will find this ‘toolbox’ approach to ethics unsatisfying, or even theoretically suspect. Mainstream moral theories are traditionally conceptualized as alternatives to each other, and their respective premises seen as making them mutually exclusive and collectively incoherent if not inconsistent. It is unclear, for example, just how Beauchamp and Childress (1994, p. 260) can argue for a form of the principle of utility “as one among a number of prima facie principles” and then argue in favour of the principle of autonomy on Kantian grounds; on the face of it, Kantian respect for persons does not admit of balancing. Of course, some version of the principle of utility may well be consistent with a principle of respect for persons, but the latter will not strictly speaking be Kantian.

So there is reason for caution in embracing this method. However, I will argue that this ‘toolbox’ approach to ethics – or some version of it at any rate – deserves our endorsement. This eclectic approach may reflect a genuine skepticism regarding the possibility of a truly comprehensive moral theory. This is certainly the case for Beauchamp and Childress (1994, p.106), who “express a constrained skepticism about…foundationalism and are doubtful that…a unified foundation for ethics is discoverable.” This skepticism seems reasonable, given the apparent failure of moral philosophers to reach anything approaching consensus at a deep theoretical level. At the very least, the toolbox approach represents an eminently reasonable stopgap measure. We don’t currently have available a well-defended, comprehensive moral theory.\(^8\) We do, however, have a short list of kinds of reasons that seem to apply particularly well to a wide range of moral problems, and which gain plausibility by virtue of their association with

---

\(^8\) As Dunfee and Donaldson (1995, p.175) note, for example, “none of the [theoretical] approaches to date have been able to establish a beachhead as a practicable, generally accepted core paradigm in business ethics.”
well worked out, if not uncontroversial, moral theories.

Some situations or issues, for example, simply seem to call for an analysis that draws heavily on the work of rights theorists. To take an example from health care ethics, questions of medical confidentiality are dominated by arguments about patient autonomy and the right to privacy, mostly to the exclusion of consequentialist arguments. The substantive conclusion most often reached here is that a competent patient’s privacy right easily overrides consequentialist considerations of whether the outcome of a particular breach of confidentiality would, on balance of probabilities, be better than the outcome associated with maintaining confidentiality. The only serious consequentialist challenge to that right comes in extreme cases, in which the health professional feels that breaching confidentiality is the only way to prevent grievous harm to some third party. In other cases, modern bioethics is decidedly consequentialist. When it comes to receiving any specific treatment (say, Cardio-Pulmonary Resuscitation), patients are not generally regarded as having a right to treatment independent of the physician’s judgment as to the appropriateness of that treatment. Rights have little role in such arguments (except in being conspicuous by their absence), and physicians are told to decide whether to offer treatments based primarily on their assessment of the outcomes associated with treatment and non-treatment respectively.\textsuperscript{9}

Given this plurality of considerations of which applied ethics avails itself, it seems natural to contend that arguments for convention-based obligations should be thought of as one more kind of consideration that can usefully be brought to bear in moral deliberation. Expansion of the set of theories to which applied ethicists appeal should not, of course, 

\textsuperscript{9} For an interesting take on the instrumental nature of applied ethics, see (Whitbeck 1996).
proceed without caution. To its credit, what I have called the ‘toolbox approach’ has been relatively parsimonious in the range of tools that it seeks to employ. Thus far, only (various forms of) utilitarianism, (more or less vaguely) Kantian theories of rights, and (usually Rawlsian) theories of justice are really pervasive. The language of virtue theory is also fairly common in applied ethics, but it gets much less attention than the other three sorts of theories already mentioned. While this list seems quite inclusive, note that it does limit the applied ethicist to appealing to historically creditable and purportedly complete moral theories. It is one thing for us to appeal to a number of well worked out moral theories; it would be quite another to appeal to a comprehensive list of human values, derived from an exhaustive accounting of cultural norms and traditions. The moral pluralism which characterizes modern applied ethics is not of the latter sort.

Note also that this toolbox approach need not be seen as a critique of theory per se. That is, it need not be based upon assumptions about the futility of theorizing about ethics. The elements of the toolbox are theory-driven. Further, the need for a toolbox (as opposed to a monistic theory) can readily be cast as itself a theoretical stance.

The pluralist approach for which I give (qualified) endorsement here is instrumental in nature. It is based on the idea that morality serves a function – roughly that it helps people ‘get along’ in social situations. We need not be dogmatic about this function. It may well be that we would have trouble establishing empirically that this is the sociobiological function which allowed moral constraint to evolve. But this is roughly the function we want

---

10 Many theorists have thought virtue theory an interesting, if incomplete, approach to ethics. Note that a toolbox approach can make use of what seems right in virtue ethics without relying on it exclusively. A similar comment might be made about Feminist ethics, which is often regarded as providing a useful critical perspective but not a full moral theory.
it to serve now. At any rate, given this goal, different social situations may require different moral mechanisms to mediate them. Thus we need a number of theoretical tools to facilitate our thinking in these areas. Rights-based approaches will work best for some issues; the establishment of a minimum wage – a legally-enforced right – has been successful in helping wage-earners avoid under-bidding each other unto poverty. Consequentialist accounts will work better in other domains; ends-based reasoning was instrumental, for example, in attempts during the 1980s to shift Canada’s drug policy away from enforcement and toward treatment and education. Of course, sometimes we will want to apply both rights-based and consequentialist thinking to a given issue, as when rights act as side-constraints on the pursuit of good consequences. The language of virtue has also had its role: appeal to virtues has arguably played a part, for example, in allowing professional organizations to socialize their members in ways that increases mutually beneficial cohesion. Similarly, I have argued in this Thesis that thinking in terms of social conventions will be useful in a wide range of circumstances (even if not convincing in all situations).

Are conventions theoretically credible enough to add to the applied ethics toolbox? We have said that the toolbox approach gains credibility from parsimony: only reasons drawn from credible moral theories are generally appealed to. Whether a convention-based approach warrants our respect is an open question; that it at least warrants serious consideration is the central contention of this Thesis. Whether or not we regard a convention-based approach as fully satisfactory, an argument for its inclusion in the applied ethics toolbox might be found in its ability to take the place there that might otherwise have been reserved for contractarianism. As noted above, conventionalism shares several
attractive features with social contract theory. In light of the near unanimous (if implicit) rejection of social contract theory by scholars working in applied ethics, a convention-based approach holds the promise of providing a moral perspective that takes seriously the idea of mutual constraint for mutual benefit (i.e., the idea of reciprocity), and that bases its claims on a theory of value shared by the social sciences. I turn now to a brief discussion of applied ethics' neglect of social contract theory and of these two aspects that conventionalism shares with that theory.

Of all major modern moral theories, social contract theory is almost certainly the most badly underrepresented in applied ethics. Beauchamp and Childress's (1994) seminal text in bioethics only mentions 'contract' in a section on Kant. Most other texts in applied ethics are not significantly different. Often social contract arguments are alluded to only obliquely in chapters on justice. Beauchamp and Childress (1994), for instance, come close to mentioning social contract theory in their discussion of egalitarian theories of justice. The example they cite of an egalitarian theory of justice is none other than that of John Rawls, best known as the leading figure in 20th Century social contract theory. Interestingly, however, the quotation (p.339-40) they choose with which to sum up Rawls's approach to justice is one in which Rawls focuses on his theory of justification (i.e., reflective equilibrium) rather than on the contractual nature of his theory of justice.

Of the various branches of applied ethics, business ethics has certainly paid the most attention to social contract theory. A review by Dunfee and Donaldson (1995) suggests that interest in contractarian approaches is growing among business ethics scholars. Further, Dunfee and Donaldson make a good case for the particular aptness of contractarian approach in business ethics. However, social contract theory is far from commonplace. For
example, Velasquez's (1998) highly regarded and representative text in business ethics discusses Rawls's account of justice as fairness, but neglects to label that account as contractarian; neither does he situate Rawls's account within the social contract tradition, nor include contract in his list (p.128) of "main kinds of basic moral considerations."

Thus it is not much of an exaggeration to say that applied ethics, for whatever reason, has ignored the social contract tradition. But in so doing, applied ethics has thrown the baby out with the bath-water. Because it does not have social contract theory to draw upon, applied ethics lacks access to a theory that takes seriously the notion that human morality is about mutual constraint for mutual advantage (i.e., the centrality of the notion of reciprocity). In failing to take seriously this aspect of morality, applied ethics risks being at odds with the respectable evolutionary perspective that says that advantage (in some sense\textsuperscript{11}) is the \textit{sine qua non} condition for the evolution of any biological phenomenon.\textsuperscript{12} Further, without this functionalist perspective, applied ethics risks being dominated by theoretical perspectives that seem at times to reverence form over function; one worries, for example, that the constant emphasis on universality is an aesthetic rather than a functional requirement. A convention-based approach can fill the gap left by social contract theory in this regard, because the former approach is grounded firmly in the idea of constraint for mutual benefit.

Further, the fact that applied ethics rejects social contract theory also means that the toolbox approach lacks access to a theoretical perspective that embodies a notion of value

\textsuperscript{11} Depending on the context, the relevant kind of advantage might be individual, reproductive, or group.
\textsuperscript{12} It might be replied that morality is a social phenomenon. However, from the sociobiological perspective, the social is a subset of the biological.
consistent with our practice in the leading social sciences. Utilitarianism, rights theories,\textsuperscript{13} Rawlsian (and most other) theories of justice, and virtue theory all make use of (though they too seldom make explicit) theories of value which accord value in an agent-independent way. That is, they attribute value – not value relative to some framework, but value \textit{simpliciter} – to pleasure, or to acting according to duty, or to leading a virtuous life. Utilitarians hold that the pleasurable experiences (or interests, etc.) of every sentient entity have value, and thus ought to be sought by all rational agents. Mill (1863, ch. iv, para. 3) claims that “each person’s happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons.”\textsuperscript{14} In modern times, leading Utilitarian Peter Singer (1993, p.101) has claimed that “the universal aspect of ethical judgments requires us to extend our positive evaluation of our own experience of...pleasures to the similar experiences of all who can experience them.”\textsuperscript{15} Likewise, moral theorist Thomas Nagel (1980) holds that the value of sensory pleasure and pain (at least) are agent-neutral, and that just as an individual’s pain is a reason for that individual to take action to eliminate that pain, so too that individual’s pain is a reason for others to want to eliminate that pain.

As suggested in Chapter 1, the standard non-contractarian theories have little to say about why it is that individuals ought to share the theories of value that these moral theories presuppose. Social contract theory does not (or at least, needs not) begin with such strong premises. While a convention-based theory is not, strictly speaking, part of social contract

\textsuperscript{13} An obvious exception here would be a contractarian theory of rights, as exemplified by (Narveson 1988).

\textsuperscript{14} As Gauthier (1986, p.52) notes, in making this claim Mill commits the fallacy of composition. Though his reasoning was fallacious, Mill’s conclusion has been popular. Gauthier notes that “…a hundred years of ever more sophisticated efforts to avoid Mill’s fallacy have not advanced the cause of utilitarianism a single centimetre.”

\textsuperscript{15} For a criticism of Singer’s theory of value, see (Danielson and MacDonald 1996).
theory, it shares with contract theory its foundation in the preferences of individuals and its emphasis on mutual benefit. Thus a convention-based approach could serve to fill the gap left by social contract theory in this regard. As noted in Chapter 1 above, we agree with David Gauthier's (1986, esp. p.46-59) excellent account of this issue. He concludes (p.59) that value is "...a measure of individual preference – subjective because it is a measure of preference and relative because it is a measure of individual preference."

I conclude that there is at least prima facie reason to include a conventionalist theoretical perspective among those upon which the pluralist approach to applied ethics draws. It has been a central contention of this Thesis that many (if not all) of our moral obligations are best regarded as conventional in nature. Even if we are not convinced that a conventionalist perspective can be considered comprehensive, a convention-based approach to ethics shares with social contract theory certain valuable components that make it a valuable element to have in the toolbox in place of the generally unpopular contractarian perspective.

We move now to our second line of defence against the criticism that a convention-based approach fails to match adequately the tenets of everyday morality.

2.0 Limited Aspirations for Moral Theory

In defence of conventions, we might also attempt a second, bolder line. Namely, we might argue that some form of conventionalism does constitute a fully adequate moral theory, given reduced (i.e., more realistically modest) expectations regarding what morality, and moral theory, can actually accomplish. The task of defending this stronger line will be broken into three tasks. First, I will argue that the demands of everyday
morality are not demands that any moral theory could, or should, meet. Second, I will point out that other factors – especially Humean moral sentiments – fill at least part of the gap left between common-sense morality and the obligations that a convention-based approach can sustain. Finally, I will argue that strong value claims upon which utilitarian and deontological theories rely have no clear basis if not grounded in conventions of some sort.

2.1 The Limits of Moral Theory

In this section, we will argue that there are significant limits on what we can – or rather should – expect from a moral theory. This argument rests on our instrumentalist approach to moral theory. That is, our approach here has been to understand ethics as playing a role, namely that of helping us to achieve Pareto superior outcomes where those are available. To ask ethics to tame constant-sum games (i.e., situations of pure conflict), for example, is to ask too much of it. As Danielson (1992, p.32) notes, “nothing can be done to civilize this type of interaction.” Morality is a tool to improve the net gain to be had from interaction. Thus the main question faced by ethics (i.e., moral theory) is how in fact to go about this task. Imposing duties of beneficence, or egalitarian norms, are two of the methods that might be used to accomplish this task. Thus any moral theory that builds beneficence or egalitarianism into its premises begs the question, by confusing mere means

---

16 Of course, ethics qua tool can also be used for other purposes. For example, the attribution of a moral duty of subservience can be used to repress an underprivileged caste. In claiming that ethics is a tool to improve social life I am really saying that this is a valued goal which ethics can help us achieve.
17 Similarly, note that some games are too easy to require ethics: at least some games of pure coordination come to mind in this regard.
18 It is of course important to see, as Danielson does, that many games of pure conflict are embedded in larger cooperative games. Thus even though only one team can win the Championship, both teams benefit from there being rules.
with ends.

Of course, real societies often do have norms that succeed in mediating constant-sum games, norms that commit people to acts of disinterested individual beneficence. Such norms do help to the extent to which agents in relevant situations see themselves as facing a dilemma; but the existence of, say, an equality norm to govern the division of cake cannot eliminate the fact that the more cake you have, the less I have. Further, in counting such norms as "moral" we are reduced to pure sociology: people are in fact sometimes motivated by such norms, but we can give no adequate account of why they should be. A convention-based account, in comparison, provides a solid theoretical grounding – based on a solid theory of value shared by the social sciences (as outlined in Chapter 1 above) – for an admittedly reduced range of moral obligations. This grounding is more solid than that of the other mainstream moral theories in that a convention-based theory need not rely on the questionable value claims upon which utilitarianism and rights theories are built.

2.2 Complementing Convention with Sentiment

In this section, we will discuss briefly the role of moral sentiments in filling at least part of the gap left between common-sense morality and the obligations that a convention-based approach can sustain. Our argument here will be closely related to that offered in the

---

19 On the face of it, egalitarian norms of justice seem also able to tame a constant-sum game, by telling the agents involved that a resource ought to be shared equally. But such games will typically be embedded in larger, non-constant-sum games.

20 A contrary view of the relative weight of common moral intuitions is provided by Thomas Nagel (1979, p.x), who writes: "I believe one should trust problems over solutions, intuitions over arguments, and pluralistic discord over systematic harmony. Simplicity and elegance are never reasons to think that a philosophical theory is true: on the contrary, they are usually grounds for thinking it false." Quoted in (Gauthier 1986, p.269 n.4).
previous section. We will argue that we should not expect a moral theory to generate obligations to do things that natural sentiment drives us (at least sometimes) to do regardless.

We noted in Chapter 6 that a convention-based account cannot readily generate obligations of beneficence. Our example was the rich man who could find no convention-based reason to give $20 to a poor man, even though the amount was trivial to the former and constituted a great boon to the latter. Perhaps more to the point, it seems that a moral theory grounded in convention could not require a transfer of even just two dollars, even if doing so would save a poor man’s life. Since a rule stipulating such behaviour would not be mutually beneficial, it could constitute a norm but not a convention.

This fact seems to make a convention-based approach a rather cold one. We generally feel that at least some kinds of disinterested beneficence are obligatory. But notice that the language I have used here is the language of sympathy. Most of us do in fact have a range of moral sentiments, sentiments that cause us to feel that we ought to help others even when there is no rational reason to do so. Nothing in what we have said here implies that we ought not to have such other-regarding sentiments. As Alan Wolfe (1989, p. 29) points out, even Adam Smith – most often cited as a defender of the power of self-interest – thought that natural sympathy would enrich and deepen the range of moral obligations in modern societies.

The conventionalist approach described here is a morality of mutual benefit. Such being the case, we have assumed that the parties to conventions have no particular interest in each other’s welfare. Gauthier calls this the assumption of mutual unconcern, or ‘non-tuism.’ But this methodological assumption in no way implies that people in fact are
mutually unconcerned. Rather, this assumption is intended to show that even people unconcerned with each other’s welfare can find rational reasons for acting morally. Gauthier (1986, p.103) notes that given such a rational justification of morality, “[o]ne is not able to escape morality by professing a lack of moral feeling or concern, or a lack of some other particular interest or attitude, because morality assumes no such affective basis.”

To begin with, we all of course have natural affections for kith and kin. The evolutionary importance of kin altruism was established by Hamilton (1964) and elaborated elegantly by Dawkins (1976). A convention-based moral theory does not (readily) imply obligations to one’s children; rather, it acts as a supplementary form of constraint, showing why it is (at least sometimes) beneficial to constrain oneself in interactions beyond the family.

But what of those for whom we have no natural affection, and with whom we have no interest in cooperation? Gauthier notes two mechanisms which may make relations in such cases less bleak than our rationality-based morality might at first seem to imply. First, Gauthier (1986, p.286) notes that we may, due to socialization, “already be disposed to” cooperate with them.21 That is, if most of our interaction is with people with regard to whom it is rational to constrain our actions, then it is likely that we will have acquired the habit – the disposition – of constraint. Those from whom we cannot benefit may themselves benefit from our psychological hysteresis. Secondly, Gauthier suggests (p.286) that we may

---

21 This and the following point is from Gauthier’s consideration of the interaction between two hypothetical nations, one civilized and technologically advanced, the other warlike and backward. Gauthier gives socialization and sympathy as reasons why the civilized nation might constrain its actions vis-à-vis the backward nation, despite apparently having nothing to gain from doing so.
feel sympathy (for whatever reason), such that we may consider ourselves obliged to render aid to someone “[w]ithout recognizing any positive obligation towards” them.\textsuperscript{22}

Harman (1993) points to three possible sources of sympathy. First, he notes (p.125) that it may be that we are “genetically constructed so as to feel such concern and respect for others....” Second, Harman suggests (p.125) that there may be a \textit{convention} of concern and respect. That is, it may be mutually beneficial for each of us to develop a habit of taking an interest in the welfare of others, and this habit may result in one taking an interest in the welfare even of those from whom one can derive no benefit. Finally, Harman (p.125-6) suggests that a sort of ‘stimulus generalization’ may occur, such that the sorts of situations that stimulate in us sympathy for kith and kin may also come to stimulate sympathy regarding outsiders, or even animals.

Thus the fact that our theory of rational morality limits our sphere of moral obligation does not mean that we will never feel obliged to render aid to those less fortunate.

\textbf{2.3 Other Principles as Conventions}

A defence of conventions as a fully adequate moral theory will need to explain the apparent salience of the principles generated by the mainstream moral theories. It may be possible to do so by integrating the norms which utilitarian and deontological theories generate. We should see this integration project as part of the argument for limiting what we expect from moral theory, since in making utility and rights (“merely”) conventional,

\textsuperscript{22} For a review of the literature on the role of empathy in influencing altruistic behaviour, see Goldman (1993).
we weaken our premises and thus weaken the moral conclusions that we can draw.

A conventionalist account, qua full-fledged moral theory, can go some distance toward assimilating the principles advocated by other, competing theories. On such an account, the Principle of Utility, the Kantian right to personal autonomy, and egalitarian conceptions of justice may themselves be conventional in nature. A conventionalist meta-ethics is happy to accord weight to some version of the principle of utility, for example, provided of course that acting on such a principle can indeed be shown to be a regularity of behaviour, sustained by an interest in coordination and an expectation that others will do their part. Indeed, a conventionalist moralist will applaud the fact that applied ethicists have reached such broad agreement on — or rather, such a pervasive convention regarding — the sorts of arguments that may be brought to bear on ethical issues.

More generally, we can imagine a moral system that demanded — as a matter of convention — that individuals not calculate their obligations solely by reference to convention. In settling the terms of their moral interaction, individuals in a society face the following dilemma:

\[\begin{array}{c|cc}
\text{Agent A} & \text{Rule 1} & \text{Rule 2} \\
\hline
\text{Rule 1} & \text{good} & \text{best} \\
\text{Rule 2} & \text{worst} & \text{bad} \\
\end{array}\]

Figure 7.1: "The Moralist's Dilemma"

Rule 1: Help others when doing so is not burdensome.
Rule 2: Help others when convention demands it.

23 Dunfee and Donaldson (1995, p.181-2) make a similar suggestion about the compatibility of contractarian thought with a pluralist approach to business ethics, as when they note that "[c]oncepts of virtue may themselves be part of an extant contract...."
It seems entirely possible that strict adherence to convention (Rule 2 in Figure 7.1) would be Pareto inferior to at least some other moral systems (e.g., Rule 1 in Figure 7.1). All would likely benefit from a system under which individuals undertook small acts of beneficence in situations in which the reciprocity relation necessary for a convention was not evident. Thus a convention could arise that required that agents not limit their moral reasoning to convention-based reasoning. Some non-conventional principles (such as some version of the Principle of Utility) could be part of a specification (either arbitrary or rational) of the set of behaviours that we might, as a matter of convention, demand of each other.

Alternatively, non-conventional principles might be endorsed on a convention-based model if they can be shown to serve some auxiliary function, such as stabilizing existing conventions or helping to select among alternative conventions. For example, it might be found to be the case that engendering a general (Kantian) respect for persons is a good way to get people to adhere to particular conventions in cases where delinquency is likely to go unnoticed. Thus a non-conventional principle might stabilize a beneficial convention that might otherwise crumble due to opportunistic defection.

Thus a convention-based account produces a lean moral theory. It is a moral theory that implies a more limited range of moral obligations, but that does so on the basis of minimally objectionable value claims. From weaker (i.e., less controversial) premises proceed stronger (i.e., more certain) conclusions. Note that Gauthier (1986, p.268) defends his own moral theory in roughly the same way: "[Minimax relative concession] gives us [all we ever wanted in a moral theory], in providing a rational foundation that does not play fast and loose with our conception of rationality; it gives us less, in providing that
foundation only within the confines of mutual benefit.”

3. Conclusion

I have offered, here, two defences against the charge that a convention-based approach fails to match adequately the tenets of everyday morality. I see no convincing reason to choose between these two defences of conventionalism: I am happy to advance both of them in parallel. On the one hand, I strongly suspect that the conventionalist approach simply is the correct meta-ethical stance. The strong value claims upon which utilitarian and deontological theories rely have no clear basis if not in conventions of some sort. On the other hand, until a (more) convincing argument to that effect can be made, I am satisfied to have added one more theoretical perspective to the applied ethics toolbox.
Chapter 8:

Clinical Standards and the Structure of Professional Obligation
In this Chapter, I propose a model of professional obligation radically different from the principle-based and virtue-based models most common in the health care ethics literature. I propose a meta-ethics of professional obligation that suggests that professional standards find their source in implicit agreement among professionals, and that this agreement has as its purpose the sustaining of the profession’s side of its contract with society.

Modern clinical practice is subject to a wide range of professional standards. Some standards – namely practice guidelines – are fairly technical standards concerning proper treatments for particular conditions. Other standards regulate the ways in which clinicians may interact with patients. An example of such a standard is the injunction against physicians having romantic relationships with their patients. Finally, there are standards that, for ethical reasons, limit the treatments that may be made available or the ways in which they are delivered. An example here would be the requirement (in some jurisdictions) that HIV-test results be given to patients in person, rather than by mail or telephone. Clearly these three are neither exclusive nor exhaustive categories, but they give us a rough idea of the sorts of standards at play in modern clinical practice. When we consider the full range of standards – from the vague and general to the detailed and specific – we see that standards governing practice are absolutely pervasive in clinical settings.

Such standards clearly put significant limits on clinicians’ autonomy. Physicians, for example, are not left to make decisions concerning the appropriateness of particular
instances or degrees of romantic contact with particular patients according to their own judgment. Such behaviour is ruled out categorically. In terms of treatment for particular conditions, a physician is simply not free to deviate significantly from the accepted standard of practice – when such standards exist – unless such deviation is done as part of a research protocol.

Given how large a role standards play in clinical decision-making, it is remarkable that discussion of the role of such standards is essentially absent from the major texts in health-care ethics. Most leading texts focus on the principles that ought to guide clinicians in their ethical decision-making, or sometimes on the sorts of virtues that they ought to seek to embody. Professional standards are seldom discussed. For example, Beauchamp and Childress, in their *Principles of Biomedical Ethics* (1994), scarcely mention standards, except to point out that ethical decision making cannot be reduced to standards of practice. Jonsen, Siegler and Winslade’s book, *Clinical Ethics: A Practical Approach to Ethical Decisions in Clinical Medicine* (1992), gives four categories of factors that must be taken into account in clinical ethical decision-making, but neglects explicit discussion of standards of practice, implicitly relegating them to their grab-bag fourth category, namely “Contextual Features.” Books such as these do on occasion refer to the guidelines of various professional associations in support of particular ethical positions, but generally do so without reflecting on the general role that such standards play in clinical practice.

Ignoring the role of professional standards in clinical decision-making has several deleterious effects. First, it pushes standard-setting off the bioethics agenda. This is problematic for two reasons: it implies that standard-setting (including the setting of practice guidelines) is not ethics, and it impoverishes bioethics as a discipline by implicitly
limiting the range of topics open to it. This is particularly true of seminal books in the field; when those books ignore issues of professional standards, the implied message is that such standards are not to be on the agenda.

Further, neglect of the role of standards in clinical decision making gives a false picture of professional autonomy. This results in an unfortunate emphasis on decision-making from first principles. Constantly reasoning from first principles is inefficient, and will often be problematic for any agent with limited time for decision-making; it will be even more problematic for physicians forced to make decisions within the time afforded by the constraints of clinical practice. Neglecting the role of standards also leaves the field of bioethics with a flawed view of professional obligation. This particular failing is the focus of the rest of this chapter.

Exactly what role do professional standards play in clinical ethical decision-making? Standards are straightforwardly the primary way in which professionals try to affect each other's actions: they are a crucial part of the self-regulation that unites all professions.¹ In order to understand this role, we need to understand the strategic nature of the decisions with which the individual clinician is faced. As noted earlier, the word 'strategic' signals the way in which various agents' actions inevitably interact. In social situations, agents do not get to choose outcomes; rather they choose actions (or "strategies"), which interact with actions taken by other agents to produce outcomes. So the strategy I should choose depends on what I expect you to do, and vice versa.

In analysing strategic interaction of this sort, we do well to begin with a thin model of the motives and incentives involved, rather than taking into account immediately

¹ For an introduction to this model of professionalism, see (Bayles, 1981).
everything that could affect the agent's behaviour. We start by analysing the motivational forces internal to the agent, rather than including in the analysis the fact that there is, for example, a law dictating that we drive on the right. If we fail to start with this thin model, it is likely to be very difficult to see how large or small a role a given institutional constraint is really playing. The simple motivational model we will use here will assume that physicians are primarily motivated by the good of their patients.\(^2\)

Let us move to an example, and investigate the situation faced by an individual physician in deciding whether or not to attempt to push aggressively to get her own patients moved to the top of a waiting list for a kidney or a private room or an MRI. A physician motivated solely by an interest in the well-being of her own patients must surely want to see her patients get to the top of such waiting lists. And physicians will often be able to make this happen, either by being the squeaky wheel, by calling in favours, or by outright misrepresentation of the patient's condition. (Of course, physicians should advocate for their patients to get better positions on waiting lists where indicated by real medical need. What is at issue in our example is simple advocacy independent of need.) But what will be the result of such well-intentioned manipulation?

If all physicians attempted to manipulate the system for the good of their own patients, the result would surely be bad for all. No one would gain any relative advantage (since everyone would be competing aggressively) and the system would surely be bogged down – rendered less efficient – by the manipulative efforts of well-meaning doctors. If, on

\(^2\) We could, of course, begin instead with the assumption that physicians are primarily self-interested. This would be closer to being the 'standard' assumption. But that assumption would be unnecessarily insulting to physicians for our purposes here; as will become clear, we can make our point with the assumption of any suitably narrow motivation.
the other hand, no physicians attempted to manipulate the system, the result would be good for all. No physician’s patients would be at any competitive disadvantage, and the system, freed of partisan meddling, would run relatively smoothly and efficiently. So clearly, it is rational for physicians as a group to prefer a system in which physicians do not engage in unwarranted manipulation of waiting lists.

But think now of the situation faced by the individual physician. If the physician in question – call her Dr. A – attempts to manipulate the system, the outcome (for her own patients) will depend on what other physicians are doing. If other physicians in general are not engaging in unwarranted manipulation, then it seems that (from the point of view of the good of her own patients) she ought to engage in manipulation. Such manipulation will result in a beneficial rise up the waiting list. She thereby serves the interests of her patients, and her unilateral meddling won’t be enough to decrease the efficiency of the system. This is a very good outcome for her patients. And if, on the other hand, other physicians in general are engaging in such manipulation, she would be a fool not to do so as well. For if she does not, her own patients will be at a relative disadvantage, and her unilateral abstention from manipulation will not be enough to make a positive difference to the functioning of the system. So it seems that no matter what other physicians are doing, Dr. A ought rationally to attempt to manipulate the system in favour of her own patients.

This strategic situation can be represented diagrammatically as follows:

---

3 Note of course that this ignores questions of distribution of benefit among Dr. A’s various patients – will the patient fought for on Tuesday benefit less than the patient fought for on Monday? I am indebted to Jonas Dow for pointing this out to me.
Dr. A chooses...

Other Physicians choose...

<table>
<thead>
<tr>
<th>Don't Manipulate the System</th>
<th>Manipulate the System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don't Manipulate the System</td>
<td>Good</td>
</tr>
<tr>
<td>Manipulate the System</td>
<td>Worst</td>
</tr>
<tr>
<td>Best</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.1: “The Doctor’s Dilemma”
Outcomes indicate result for patients.

Note further that the situation is symmetrical. All physicians are in the same situation, so all will face the same temptations as Dr. A. Therefore all physicians will be motivated to manipulate the system, regardless of what other physicians do. The result we should expect, then, is a situation in which all physicians manipulate the system, despite the fact that (as I argued above) it is rational for physicians as a group to prefer a system in which they do not engage in unwarranted manipulation of waiting lists. If this bad outcome comes about, it will be as a result of individual physicians pursuing the good of their own patients.

But in fact, the situation predicted by this model does not generally obtain. Manipulation of waiting lists is not rampant, as would be expected if physicians were motivated only by the good of their patients. Rather, physicians follow a professional standard (mostly unspoken) that says that they ought only to agitate for the legitimate needs of their patients. This standard serves as an additional motivating factor on top of physicians’ desire to do what is best for their patients. It is the existence of this standard...
that explains why in fact physicians, despite being motivated by the good of their patients, do not end up in the bad situation that we are driven, on the basis of our initial simplistic motivational assumption, to predict.

Lest we should think that the analysis provided here is unique to the question of manipulation of waiting lists, we can point out other decisions which physicians must make and which share the strategic structure illustrated in Figure 8.1 above. We will see this structure, for example, in the physician’s decision whether or not to misrepresent a patient’s condition on an insurance form; in the physician’s decision whether to give antibiotics that may not be strictly necessary (thereby diminishing the long-term usefulness of that antibiotic in the population); and in the physician’s decision whether to order extra diagnostic tests that are of minimal use and that would have to be paid for by the provincial health programme. In each case, the well-intentioned physician may be tempted to do what is best for her own patient. And in each case the net effect of all physicians doing so would be disastrous.4

We thus have an explanation of the source of physicians’ obligations that conflicts with the more standard models. The standard bioethics model seeks to define professional obligations in terms of principle-driven obligations to patients (however defined) or to society as a whole. The model proposed here instead describes professional obligations in terms of duties imposed by the profession upon its members (or rather by the members upon each other) and enforced for their common benefit. This model sees moral obligations

4 To emphasise the generality of this model, I should note that it also covers standards regarding confidentiality, informed consent, romantic relationships with patients, and so on. In such cases, however, professional standards serve to restrict physicians’ pursuit of their narrow self-interest, for the good of the whole profession.
as having their origin in social convention. That is, I argue that this particular set of moral obligations – those embodied in professional obligations – can best be seen as a set of regularities of behaviour, sustained by an interest in coordination and an expectation that others will do their part.

If obligations are seen as serving the function attributed to them here – the constraining of behaviour, for mutual benefit – then obligations will only be possible to the extent to which the agents involved have something to gain from each other’s exhibiting constraint. If Dr. A did not care whether other physicians limited their manipulation of waiting lists, she would have little reason not to engage in such manipulation herself. This perspective casts doubts upon the extent to which physicians can be seen as having direct obligations to their patients. Given the notorious power imbalances inherent in the doctor/patient relationship, such relationships are deficient (though not entirely lacking) in one ingredient – namely an interest in coordination – that is necessary for the existence of an obligation-generating convention. Physicians have little to gain directly from moral agreement with particular patients. If moral obligations have their source in mutual constraint for mutual benefit, any attempt to describe the obligations which physicians have in terms of their relationships to patients will surely be deficient. A physician’s professional obligations, then, will primarily be obligations owed to other physicians.

Some will be alarmed or even shocked at the idea of a model of professional obligation based primarily on the needs of the profession. At some level, this worry may be warranted: as consumers of health-care services, it would perhaps be better for us if all
physicians saw themselves as subject to clear and direct moral obligations to us. But two points should be made that may temper that concern.

First, we should recall that nothing in this perspective assumes that physicians feel or ought to feel no direct obligations to their patients or to society as a whole. Clearly they do feel and have such obligations. Physicians have certain direct obligations to patients, obligations that have nothing to do with membership in a profession. Physicians have, for example, as a result of the general social convention condemning lying, an obligation not to lie to patients in any but exceptional circumstances. This will be true for a variety of obligations which we normally see ourselves as owing to each other as fellow citizens. But none of these will count as professional obligations.

Second, notice that on the model given here, the way in which professionals seek their own ends is precisely by adhering to such a high standard – by responding to our needs so well – that we see them as worthy of our trust. To say that professional obligations are enforced as a way of meeting the needs of professionals is less worrisome given that the needs of professionals so clearly include a need to do good for their patients. The profession’s need for cohesion here stems from the contract with society to which professions are typically seen as being party. According to this (usually implicit) contract, physicians constrain their practice by adhering to certain standards (both clinical and ethical); in return, physicians as a profession are granted a monopoly on a wide range of

---

5 A closer look reveals that this is not the case. For recall that, as illustrated above, what is individually rational can be collectively disastrous. While it would be individually rational for each of us, as a patient, to enjoin our own physicians to serve our own individual good, the collective result (were our physicians to comply with such requests) would be bad for all of us, as shown in Figure 8.1.
medical services, a high degree of professional autonomy, good financial remuneration, and 
prestige. This contract situates us in the upper left quadrant of Figure 8.2.

<table>
<thead>
<tr>
<th>Give Docs Autonomy &amp; Prestige</th>
<th>Give Socially Best Care</th>
<th>Follow Own Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Good</td>
<td>Best</td>
</tr>
<tr>
<td>Worst</td>
<td>Worst</td>
<td>Bad</td>
</tr>
<tr>
<td>Best</td>
<td>Bad</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.2: “Contract Between Society and Physicians”

In order to gain the advantages of this bargain with society, physicians as a group must provide the kind of care that society demands. And what society demands is for physicians to give good care, but also to do so with an eye to broader issues of community health and the functioning of the health care system. This implies abstention from undue manipulation of waiting lists, from giving marginally necessary antibiotics, from ordering extra diagnostic tests, etc., even when such activities seem to be in the best interests of particular patients. The profession as a group must find a way of providing the good demanded by society. Providing that good requires cohesive action, and the requisite level of cohesion comes through professional standards and the obligations they imply.

Note that if we, as a society, want to demand certain behaviour of physicians, it will not do simply to assert that individual physicians have obligations not just to their own patients but to society as a whole. The individual physician, caught in the dilemma represented by Figure 8.1, cannot make the system efficient on her own. Even if she feels a general obligation to society, and so abstains from unnecessary manipulation of the system,
enough other physicians manipulating the system will nonetheless render it inefficient. If
she abstains in such a situation, she will have sacrificed the good of her own patients for
nothing. Collective action is required.

I have proposed here a meta-ethics of professional obligation that suggests that
professional standards find their source in implicit agreement among professionals, and that
this agreement has as its purpose the sustaining of the profession's end of its contract with
society. This model provides a much-needed grounding to discussions of the limits of
professional obligation and professional autonomy. Such a model also takes seriously,
without endorsing, the power imbalance inherent in the physician-patient relationship. This
implicit agreement constitutes a convention of the sort we have been interested in
throughout this Thesis.
Chapter 9:

Conclusion
Humans "cannot change their natures. All they can do is change their situation."
– David Hume, *Treatise*, Book 3, Part II, Sec. VII (pg. 537)

This Thesis has been about the role which social conventions play in shaping our moral choices, and about the possibility of a normative theory that takes such conventions seriously. It also hints at the idea of looking at conventions as a kind of *moral technology*. If existing conventions serve a useful function, then perhaps we can take what we have learned about conventions and apply that knowledge in a forward-looking manner. Let us briefly recap the points covered in this Thesis, before saying more about where it has led us.

In Chapter 1, we outlined the shape of this project, and explained its roots in methodological individualism and a relative, subjective theory of value. Chapter 2 surveyed the literature on norms and conventions, and explained why it is that despite the prevalence of the former in the literature, moral theorists should focus on the latter. Chapter 3 looked at the ways in which conventions structure strategic interaction. We found that it did so both by providing information that served as an input to rational decision making, as well as by providing a more direct, non-instrumental form of motivation. In Chapter 4, we looked at the relevance of the literature on public goods and the problem of collective action, and argued that beginning and maintaining conventions embodies a collective action problem. In Chapter 5, we moved to examine the normative force of conventions, and suggested that conventions constitute a reasonable set of constraints on self-interested behaviour. Chapter 6 addressed the fact that a convention-based approach seems capable of endorsing asymmetrical outcomes, at least some of which seem unjust. In Chapter 7, we looked more
generally at the adequacy of a conventionalist account, and examined the departures which such an account seems to make from everyday morality. Finally, in Chapter 8 we put theory to practice, and examined the world of professional obligation. We offered there a meta-ethics of professional obligation that suggests that such obligations are best understood as conventions between professionals. That argument serves as a challenge to existing principle-based and virtue-based theories of professional obligation, and illustrates the practical importance of this Thesis.

Next, can we take what we have learned about conventions and apply that knowledge in a forward-looking manner? That is, does what we have learned here offer us guidance when it comes to changing existing conventions and building new ones?

**Changing Existing Conventions**

Robert Sugden (1990, p.781) argues that “Given an instrumental view of rationality, it will rarely be rational for an individual to challenge a well-established convention.” Sugden is thinking, of course, primarily of conventions to solve coordination problems. Situations involving mixed motives clearly may provide an agent with grounds to challenge existing patterns of interaction; in a mixed-motive game, any agent will be tempted to challenge the convention if she thinks she can get away with unilateral deviation. Further, it may be instrumentally rational for an individual to challenge even a well-established coordinating convention if she can point to a potential convention that is both Pareto-superior and easy to implement.

Sugden (1990) argues that the most important aspect of conventions is that they provide a common body of experience which each of us can draw on in choosing actions in
situations requiring coordination. If you and I both have the same convention to refer to, that may give salience to one of several possible equilibria. According to Sugden, asking you to deviate from a convention (e.g., “Drive on the left this time!”) means asking you to ignore the epistemic advantage which the existence of a convention implies. Sugden (p.782) writes: “If I am to make any progress, I must convince you that our common experience is irrelevant, that somehow the game we are playing now is different from all those other games we have played in the past.” But there is another possibility. Even the games of pure coordination with which Sugden is concerned can have better and worse equilibria. Thus in order to change conventions, I may try to persuade you that this particular regularity is not only arbitrary, but sub-optimal, i.e., that there is accessible to us a Pareto-superior way of interacting. And I may go so far as to promise — or threaten — to deviate unilaterally.

Sugden (1990, p.782) further claims that “[i]n trying to convince you to ignore your experience, I am trying to persuade you to treat your game with me as a one-off interaction.” Sugden has in mind here social interaction that is essentially anonymous due to the large number of participants, so that even though embedded in a larger social context, two agents can plausibly treat their interaction as isolated. But even given this qualification, it is not necessarily true that, in attempting to persuade you to change conventions, I am trying to get you to see this as a one-off interaction. Indeed, were I to succeed in getting you to see this as a one-off interaction, I could well be undermining your motivation to cooperate. More often, rather, I will want you to keep in mind the fact that this interaction is part of a larger string of interactions; but in so doing, I will be trying to persuade you to be forward-looking, to participate in the establishment of a new pattern of interaction. Our
interaction will, of course, be like a one-off interaction in that neither of us will have access to a range of experiences to look back upon. However, our interaction may be unlike a true one-off interaction to the extent that our expectations about the future interactions may colour our choices.

Sugden is pessimistic about trying to convince another agent to ignore previous experience, since many one-off games have no uniquely rational solution. But note that one need not find a uniquely rational solution. If an existing convention can be shown to be Pareto sub-optimal, then all we have to prove to motivate change is that a different convention would be Pareto superior.

**Building New Conventions**

One of the implications of this Thesis is that good conventions make good moral choices easier. It might even be said that the 'real action' of morality goes on long before the individual is ever faced with a decision. This points to a need for attention to the building, rather than the following, of moral rules. Is it possible to engineer new conventions? Are there some kinds of conventions which are impossible to build and maintain, and which therefore ought not to be sought after? In earlier Chapters, we described various characteristics of conventions. Here, we seek to use that earlier analysis to determine what characteristics successful conventions require. If conventionalism is to be a truly practical theory, it should have something to say about how we can go about using this moral technology to solve moral problems. That is, an understanding of conventions should give us the tools to effect positive change in our world. While it may be true that the origins of many conventions are shrouded in mystery, and buried deep in our
social prehistory, this does not imply that new conventions cannot be started.¹ In fact, new conventions arise constantly. And some of the characteristics pointed to in our positive theory of conventions suggest that conventions can be built intentionally.

Many social problems are the result of failures of social coordination. Solutions to such problems will require the creation of new conventions.² An example cited by Brown (1995, p.45 ff.) is the nascent convention requiring households to recycle. Given that recycling can only have real impact if broadly implemented, failure to recycle is a failure of social coordination. According to Brown, a new convention, under which individuals see themselves as having an interest in the collective goal, and expect their neighbours to comply, could solve this problem. In this section, we re-examine the characteristics which conventions normally have, in an effort to reverse-engineer a template for solutions to problems of social coordination.

Let us examine a problem in need of a convention. Rather than imagine a hypothetical problem, let us work with a real one. According to Ross (1998), paired kidney exchanges³ are currently hampered by the apparent need to conduct simultaneous surgeries. Simultaneity is deemed necessary in order to preclude one donor backing out halfway through the exchange (i.e., after his or her spouse has already gained a kidney). Modern bioethics has a strong tradition of honouring the patient/participant's right to refuse any and

¹ DeBeers, the company with a virtual monopoly on the world's diamond supply, is said to have started the tradition of giving a diamond ring as a token of engagement. They are currently in the process of trying to start a new tradition around the concept of a '10-year anniversary ring.'
² See Brown (1995) for more on this theme.
³ A 'paired kidney exchange' transplant goes like this. Patient A's spouse is not a compatible donor for Patient A. Patient B's spouse is not a compatible donor for Patient B. But Patient A's spouse is a compatible donor for Patient B, and Patient B's spouse is a compatible donor for Patient A. The two pairs thus swap kidneys. Each patient therefore gets a kidney, not from their own spouse, but as a result of their spouse's donation.
all medical interventions, so it cannot require donors to go carry out their commitments. But simultaneity is problematic. The fact that relatively few hospitals have the facilities to carry out two kidney transplant operations at once means that the number of places where paired exchanges can be done is greatly reduced. This not only has bad consequences for at least some potential kidney recipients, it also poses problems of distributive justice given that suitable hospitals will be found in only the largest urban centres. This need for simultaneity would be eliminated if there were in effect a convention making it unconscionable for a donor to back out once the process is underway.4

We have defined a convention as a regularity of behaviour, sustained by an interest in coordination and an expectation that others will do their part. Let us look at each of those characteristics in turn.

One of the problems in establishing a new convention will be the question of multiple equilibria. In most cases, we will be able to imagine a range of conventions that would ameliorate our situation. Thus even given a desire to cooperate, we may be at a loss to figure out how to go about doing so. Here we can look to Sugden’s (1989) idea of families of conventions, or what we called in Chapter 2 a ‘metaconvention.’ Recall that Sugden (1989, p.94) argued that the similarity which a novel situation, “A,” bears to a familiar one, “B,” may lead to the establishment of a convention to govern A bearing a family-resemblance to the convention governing B. Sugden offered this as a possible explanation of how equilibria do in fact get chosen. Yet if we have the ability to reason

4 Thanks to Peter Danielson for pointing out how the existing moral regime in this area is Pareto sub-optimal.
analogically in this way, this suggests a strategy for facilitating the establishment of new conventions. If we want a new convention to take root, we would do well to use a convention that constitutes a strong analogy to a well-known convention in another domain, and make sure that the proposed parties to the new convention see the analogy clearly. Thus in our kidney donation problem, we might propose analogies to other sorts of promises which are generally thought to be strongly binding independent of their immediate personal consequences.

Next, if conventions are generally sustained in part by an interest in coordination, this suggests that we ought to make clear to people the fact that they benefit not just from the existence of kidney exchange programmes, but from the existence of a related convention of commitment. We ought to point out to parties to paired kidney exchanges that it is only because of the understood obligation to follow through that such exchanges are even possible.

Finally, if conventions are also sustained in part by the expectation that others will do their part, we ought to make this clear to participants in paired exchanges as well. The mutual expectation might even be made more clear by introducing the concerned parties to each other. (Under the current regime, the exchange is carried out without the parties knowing each other’s identities. In fact, hospital staff go to great lengths to keep this a secret. This is presumably to ensure that participation is fully voluntary, e.g., to preclude moralized constraint.) The role that expectations of compliance plays in maintaining

5 For an argument for the special appropriateness of analogical reasoning in political philosophy, see Aronovitch (1997). Aronovitch contends (p.78) that analogical reasoning in politics "involves building innovatively upon the past by constructively extending precedent," such that "stability, agreement, and progressive steps into the future are secured."
conventions also suggests that we ought not to be reticent about moralizing in appropriate situations.

In some situations, the appropriate vector or method of transmission of the new convention will be plain. Perhaps the easiest case is that suggested by Chapter 8 above. Professions at least sometimes have their conventions codified into codes of ethics, practice guidelines, and so on. Thus if we want to add a new professional obligation — a new convention — to the existing set governing a profession, we can begin by entrenching it in the appropriate written documents. Our attempt to build a new convention governing kidney exchanges might best start where the method of transmission is most obvious. Given that the key moral standards governing paired kidney exchanges are professional standards of voluntariness of treatment, we ought to begin by attempting to change the way medical professionals see this issue.6

Note that building and maintaining new conventions will involve instilling in ourselves new (or at least repurposed) dispositions to comply. Recall the quote from David Hume at the head of this Chapter. Hume expresses doubt that humans can change their nature. In Chapter 1 we echoed this doubt, and proposed looking instead to artifices such as conventions. Yet if conventions qua social engineering actually change our dispositions, then perhaps this pessimism was hasty. For perhaps in changing their situation people can affect the habits they acquire, thus indirectly changing their nature.

6 Note further that both Axelrod (1997) and Hirschleifer (1995) point out the importance of influential individuals in getting new regularities of behaviour started. In medical settings, physicians fit the bill.
Appendix 1:

Games
Here are the basic games referred to in this Thesis. The numbers represent agents’ ordinal rankings of outcomes, with zero being worst and higher numbers being better.

<table>
<thead>
<tr>
<th>Agent 2</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The Prisoner’s Dilemma

<table>
<thead>
<tr>
<th>Agent 2</th>
<th>Do X</th>
<th>Do Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do X</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Do Y</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

The Battle of the Sexes
<table>
<thead>
<tr>
<th></th>
<th>Swerve</th>
<th>Straight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swerve</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Straight</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Chicken
Bibliography


