Limited Conventions about Morals

Marinus Ferreira

A thesis submitted in fulfilment of the requirements for the degree of Doctor in Philosophy

in Philosophy, the University of Auckland, 2017.

Abstract:

In this thesis I describe how conventions specify how to put normative principles into practice. I identify a class of recurring situations where there are some given normative principles in effect, but they underdetermine what each individual should do, and what is best for an individual depends on what others do. I demonstrate that in such cases, whenever the community develops a response that repeatedly brings them to as good an outcome as is available according to their principles, that response is a Lewisian convention where the benefit of an outcome to each individual is measured by the extent to which it conforms to the principles they subscribe to. Since these conventions are constrained by the normative principles, I call them *limited conventions*. They are supplements to the principles, and are ineradicably involved in moral action insofar as the abovementioned cases of moral underdetermination are in play. That has the consequence that in these cases the only reliable way to follow your principles is to follow the relevant conventions. As examples of this mechanism I offer a conventionalist analysis of authority, such that the commands of an authority is normative when they instantiate a limited convention, and of the variation in understandings of virtue and vice across societies, such that the evaluative vocabulary of each society is a set of different limited conventions about how to express in word and deed the evaluative points of the virtues and vices in question. Finally, I discuss how conventions and similar forms of guidance provide a way for individuals to participate in their community's moral life without having a full understanding of the principles that underlie it, or even if they are profoundly ignorant or outright mistaken about the demands of morality.

For Elspeth

Acknowledgements:

I wish to thank my supervisor Fred Kroon for his exemplary support and feedback. I also wish to thank my co-supervisor Glen Pettigrove for his many useful comments and pieces of advice. I also wish to thank Rosalind Hursthouse, Denis Robinson, Patrick Girard, and Justine Kingsbury for the great amount of mentoring, feedback, and advice they have given me along the way.

I have benefited greatly from the Normative Philosophy Reading Group at Auckland, and the input from Glen Pettigrove, Christine Swanton, Tim Dare, Monique Jonas, Arie Rosen, Sarah Anderson, Matt Pettyman, Mark Tan, Warren Nock, and Marco Grix. I wish to thank Emily Park, Brett Calcott, Matheson Russell, and Patrick Girard for organising a writing retreat in 2016 and for their feedback and advice on that trip. I also wish to thank the audiences at various iterations of the New Zealand Association of Philosophy conferences and the Auckland Philosophy Graduate Seminar series.

I wish to thank the people who helped proofread this dissertation: Elspeth Hocking, Rosalind Hursthouse, Matteo Ravasi, Grace Alty, and Warren Nock,

I am grateful to have had my study funded by the University of Auckland Doctoral Scholarship, and for the employment I received from the Discipline Area of Philosophy.

Finally, and most importantly, I wish to thank Elspeth Hocking, who has given me wonderful support throughout the entire project, and without whom this would not have been possible.

Table of Contents

Introduction

The problem I deal with in this thesis is that sometimes we only receive incomplete guidance when trying to determine how to respond to some situation. It is possible for us to have guiding principles that tell us a lot about what we should do, and which we are happy to subscribe to, which nonetheless can't tell us everything we need to know in order to make a good decision. This can happen when the principles tell us what options are unacceptable without identifying a uniquely best option, or when the same end can equally well be reached by a range of different and incompatible ends. As problematic as this on its own would be—and people often view uncertainty as a problem in its own right—this difficulty becomes especially pronounced when we are in a situation where it's not just what we ourselves should do which is at stake, but also what the people around us will do. My suggestion is that we handle this by way of conventions that have been established in order to supplement the principles. I call my development of the problem case the *strategic underdetermination problem*, and in this thesis I argue that what I call *limited conventions* are the best way, perhaps the only way, to reliably overcome this problem in recurring situations.

This introductory chapter is devoted to scene-setting and throat-clearing. I introduce the two main features of my account, the strategic underdetermination problem and limited conventions. I also give an overview of the work done in the thesis, and what it offers to the intended audience.

I. Introducing the strategic underdetermination problem

A full discussion and defence of how limited conventions address strategic underdetermination problem cases will have to wait till Chapter 1, but I will here introduce and motivate the problem.

In very many instances we get over uncertainty about what other people will do by

determining what they should do by the lights of some normative framework. A good example of this approach is its prominent place in the rational-choice theoretic framework, when we often decide what to do by assuming the other parties in the situation follow some given standard of rationality, deducing what they would do if they were rational in that way, and then choosing the action that has the best pay-off when they've acted in that way. But if that framework only provides incomplete guidance, then there are multiple possible things the other people could justifiably do by the lights of that framework. This means I don't know what they will do, which in these problem cases mean that I don't know what I should do either. This happens even if I have some method of my own for handling the incomplete guidance, such as choosing one option willy-nilly, or not stopping to consider all the options and doing the first sufficiently good one that comes to mind. Without having some shared framework for handling incomplete guidance, the uncertainty that arises is something like a contagion that undermines the ability of anybody to reason towards their desired ends in one of the many situations where what we should do depends on what other people do. This thesis is a sustained treatment of that problem, where I show how conventions of a particular kind are the solution, and then survey what moral reasoning and action-guidance in general is like when such conventions are involved.

Often, for us to be effective in pursuing our ends, we need to depend on those around us acting in certain ways. This is because often our actions will only have the desired effect if the people around us in turn act in a certain way. For instance, it is effective for me to wait in a queue to get served in most stores but often not at bars. My ability to get served depends not just on what I do, but also on what the server (cashier or barkeep) does, and what the other patrons do. If my way of approaching the server doesn't get me served, then my planned action won't be effective. In most stores standing in a queue is an effective way to get served, because it is the shared expectation of the server and myself (and the other patrons) that people who

wish to be served join the queue in front of the server and get served when they reach the front of the queue. This means that all the patrons who want to be served join the queue, and people outside the queue are understood to not want to be served at that moment. This means that my standing in a queue is likely to be effective in getting served because it means other patrons are unlikely to work at cross-purposes with me (by barging in front of me) and the server is likely to serve me in the expected manner once I get to the front of the queue. This doesn't work in many bars, because the queueing arrangement isn't in effect. It doesn't matter whether this is because bars are often too cramped to allow queues, or there is the practice of staying at the bar counter after you've been served (meaning others don't move forward into your place), or for whatever other reason. In a bar it is often the case that other patrons don't join a queue behind you if they want to be served and you're already waiting; instead, everybody who wants to be served makes their way to the counter as directly as they can, and try to attract the barkeep's attention once they're at the bar. This means that queueing in a bar is at best a sub-optimal way to get served; if the bar is busy enough, you may never get served as people keep barging in front of you while you're waiting for others to move out of your way as they would if you were queueing. Similarly, if you engaged in this same make-your-way-to-the-front behaviour in most stores, you would be flouting the shared expectations about how people act when they want to be served, frustrating other people's attempts to get served, and are likely to be the object of disapprobation. If you didn't know the shared expectations, this difference in behaviour amongst patrons and servers would be mysterious, and you would struggle in your efforts to be served. Uncertainty about the actions of any of the parties involved then becomes uncertainty about what I should do, because it is uncertainty about the effects of my actions. That is the strategic underdetermination problem, which I call the 'SUP' for short, and instances of it 'SUP cases'.

In short, a SUP case is one that requires an agent to choose a course of action, and where

which action is best for them to pursue depends on what the other parties in the situation do, but where the principles you would use to predict their actions underdetermine what they should do. It is strategic in the rational-choice-theoretic sense in that the outcome of your action depends not only on what you do but also on what the people around you do. And in the strategic cases in question, your best resource for predicting what course of action your fellows will adopt—deducing what they will do based on the guiding principles you know they subscribe to—is undermined by the fact that those principles don't give you enough information for you to make the necessary predictions. The queueing-case discussed above is an example, because taking just the facts that there is a group of patrons who are trying to get served by a server, it is underdetermined whether the patrons queue or don't; queueing is only reliably effective if everybody expects to do so and does in general do so, whereas queueing if it isn't the usual thing is likely to be ineffective. So, we have a strategic case (getting served efficiently depends on what others do in addition to what you do) which is subject to underdetermination (whether to queue or not).

It is important to stress that the only way in which underdetermination features in my account is by way of strategic underdetermination. In the thesis, whenever I introduce underdetermination, it is as a means to establish strategic underdetermination. Underdetermination in its own right isn't pertinent to my purposes, since in cases of non-strategic underdetermination your ability to navigate through the situation depends only on yourself. In contrast, in the strategic case individuals are in relationships of mutual dependence, and I seek to demonstrate how these relationships combine with the normative force of the underlying principles in order to generate binding obligations.

Underdetermination comes in many forms. To illustrate the kind of thing I have in mind, I'll give examples of what I take to be the two most common kinds of underdetermination: where there are two determinate options available but no way to determinately choose between

them; and where there aren't determinate choices available but only vague ones.

The first kind of underdetermination is Buridan's Ass cases, where deliberation comes to a point where there are multiple options available that are clear in themselves, but it isn't clear which should be preferred over the other. Buridan's Ass is caught in indecision between two different piles of hay, both of which would be equally good to eat and is equally easy to get to, but doesn't commit to one pile or the other for want of knowing which is best. In a Buridan's Ass case there is no doubt about how to go about pursuing one of the options, but there is no determinate way to choose among the options. Other examples of this kind are deciding which side of the road to drive on, or whether to report temperatures in degrees Fahrenheit or Celsius, and so on.

The second kind of underdetermination is that which arises when our principles come to vague conclusions. In the vagueness case our deliberation reaches a point where it settles on a single option, but it is unclear what would count as pursuing that option. A familiar example of this in strategic cases is where a group of people need to settle on a bright line standard in a domain with lots of variation, such as what counts as the age of majority in a jurisdiction. In different jurisdictions the age of majority ranges from 15 through to 21, indicating that there is very widespread agreement about what the age range is where someone becomes mature enough to take on the responsibilities of adulthood, some point in their late teens or early 20s. However, there doesn't seem to be a firm answer to what the correct age is, because different individuals become mature at different points in their life, and because what is expected of mature adults is different in different societies, and so on, which also affects when you'd consider someone to have met that standard. So, we have some guidance about where to draw the line, but only vague guidance.

As commonplace and familiar as instances of vagueness are, I use Buridan's Ass for my paradigm throughout the thesis. This is because for my purposes we can easily translate any

case of vagueness into a Buridan's Ass case. To do so, we only need to cast the different available precisifications of a vague term as the different equally-attractive options that we don't have a determinate way to choose among. Then the question of where to draw a bright line through vague territory has the same form as the question of what alternative out of an underdetermined range to select. There are of course serious philosophic issues that arise when we try to do precisifications of vague terms—higher-order vagueness, for instance. My point isn't to sweep these under the carpet, but to point out that whatever turns out to be the right way to draw up precisifications, and whatever the range of available precisifications are for a vague term, we can treat it the same as a Buridan's Ass kind of case. And I do so throughout the thesis.

Let's return to the queueing example from earlier in this section just to reiterate how Buridan's Ass cases become instances of strategic underdetermination. When I enter an establishment where people approach and are helped by service-staff placed behind a counter, I need to ascertain what the best way is to attract the service-staff's attention. Merely the fact that I need to approach the counter doesn't settle whether they should do so in a single queue, or to make my way to the counter as best I can. Thus far the issue is one merely of underdetermination. But whether the course of action I choose turns out to be effective depends not just on my decision, but also the decision of the other patrons of the establishment. If I'm at a bar and they are all pressing their way to the counter while I try to queue, my action will be Quixotic and ineffective. If they queue and I try to make my way to the bar, my action will be so rude (by way of frustrating their efforts to be served) as to be counterproductive. This means I have an interest in slotting into whatever the established practice at that establishment is, which both solves my personal uncertainty and the strategic underdetermination faced by all the patrons.

Of course, the SUP doesn't exhaust all the ways in which it can be hard to predict how

other people will try to follow their principles; it is worthwhile to distinguish the SUP from them. For one thing, principles are normative rather than straightforwardly descriptive. This means they allow for things such as someone who is in general committed to telling the truth nonetheless lying to stave off embarrassment. The extent to which someone fails to live up to the demands on their principles undermines your ability to predict their action on that basis. Similarly, even someone who is sincere and conscientious in their application of the principles may fail to determine what their principles tell them to do in a given situation, either because they don't know relevant features of their situation or fail to make the right deduction. For an instance of the first kind, someone who wants to please all their guests at dinner may fail to realise that one guest is a coeliac and won't be able to enjoy the freshly baked bread; for an instance of the second kind, someone may try to make an investment with an aim to maximising long-term returns but make a sub-optimal investment because they don't realise the extent of gains to be made with compound interest. Failing to keep to your principles or failing to realise what they recommend in a particular case are ever present worries, but the SUP can arise even when these worries are averted. If the principles are genuinely underdetermining, no amount of conscientiousness or insight will suffice for them to guide you to a uniquely right response. And just as they would fail to tell you what you should do, so too they will fail to let you predict what your fellows would do.

II. Introducing limited conventions

The solution I offer to the SUP is an extension of Lewisian conventions I call *limited conventions*. What Lewisian conventions bring to the table is that they allow us to analyse the underdetermined choice between different outcomes as a choice between different co-ordination equilibria. These are the outcomes that are preferred by every participant to any other outcome that would result if a single participant (either themselves or someone else) were to change their strategy, and hence outcomes that individuals have no incentive to deviate

from.[1] Lewis then goes on to give an analysis of various social phenomena as people regularly conforming to some structure of expectations such that they reliably come to an equilibrium outcome, and have no weighty reason not to conform. Understanding strategic underdetermination problem cases as the choice between different co-ordination equilibria means that any settled choice of one option over another will be a Lewisian convention, if what we mean by 'settled' is that it is an object of the kind of structure of expectations that Lewis calls 'common knowledge'.

The limit in 'limited conventions' comes from how the range of equilibrium outcomes are limited by the background principles. They are Lewisian conventions where the extent to which an individual prefers the outcomes in question is the extent to which it conforms to the principles they subscribe to. This could be a matter of degree, such that we can have a fine-grained ranking of outcomes based on their value by the light of the principles, or it could be an all-or-nothing matter, where the principles divide all outcomes into a class of conforming outcomes and a class of non-conforming ones and any conforming outcome is preferred over a non-conforming one. What matters is that it is by the lights of the principles that outcomes are seen to be attractive or not. This limits the scope both of what can make some outcome preferable to another and of the conventions that can arise, hence the name 'limited conventions'.[2]

My analysis is more robust than the usual ways of applying conventions to the moral case, because it is more modest. There has been the occasional attempt to argue that our moral

[1] They are thus a stronger version of Nash equilibria: for a Nash equilibrium it only needs to be that the outcome in question is preferable to each agent to any other outcome that could be reached by a change in behaviour by that same agent.

[2] Lewis has no such limitation in his work, which arises from him not providing a substantial theory of what preferences amounts to. His lack of commitment to a particular theory comes into play when discussing the way in which a Hobbesian social contract may under one description fit your preferences and under a different description may not. David Lewis, *Convention: A Philosophical Study* (Malden, MA: Harvard University Press, 1969). 93-95.

principles are the result of conventions or some other strategic arrangement.[3] This thesis is not an example of that approach. What I do here instead is take the existence of guiding principles as a given, and show that conventions have an important role to play in deriving determinate action-guidance from these principles. So, on my analysis conventions supplement principles rather than being their origin. As I go on to show, they are an ineliminable part of any determinate action-guidance we have in cases where our principles are vulnerable to the SUP.

A sketch of my argument for limited conventions goes as follows: if general principles don't settle every moral question, then SUP cases arise; what is at issue with these cases is how to apply the principles; SUP solutions specify in particular cases how to apply those principles, meaning they specify a choice between different co-ordination equilibria; on my analysis people conform and should conform to the specified option because of a structure of expectations that effect; thus, any SUP solution is going to count as a Lewisian convention; thus, conventions are part of what determines the application of principles.

The reader may wonder why I am using Lewis's analysis of conventions, since in the time since his work a wide range of competing analyses have arisen. Some of these are likely to serve at least as well as Lewisian conventions. There is no deep philosophical reason why I have stuck with Lewisian conventions, but it has a number of convenient features. It has the small philosophic advantage that the SUP is the exact analogue of the coordination problem Lewis addresses, as I discuss in Chapter I, §1. But the main reason is that Lewis's analysis is the touchstone in the field, and I exploit the fact that my reader is likely to be familiar with the analysis as well as harness the very large literature that has arisen around it.

Something which does warrant a mention is evolutionary game theory. When I get to the details of my account, and especially when I talk about the epistemic role of conventions in

[3] See the survey in H. Peyton Young, "The Evolution of Social Norms," *Annual Review of Economics* 7, no. 1 (2015).

decision-making, the reader may worry whether I wouldn't be significantly better served by using evolutionary game theoretic analyses of conventions, such as the extensive work of Brian Skyrms translating the Lewisian analysis into an evolutionary model.[4] Whereas Lewis puts in place strict epistemic requirements, such as his standard of 'common knowledge', it is commonplace in evolutionary game theory for the individual agents to have a limited appreciation of the situation they find themselves in—at the limit, Skyrms and others show how even creatures that plausibly have no mental states, like insects and even bacteria, can participate in the evolutionary analogue of conventions. Since the one point on which I put stress on Lewis's account is on the epistemology, it seems I've missed an opportunity to use an analysis that is better suited to what I'm doing.

The SUP can arise in any domain where action is regulated by general principles, but I concentrate on moral cases because that is where conventions are the hardest to establish, and, I believe, the most interesting. Throughout the thesis I discuss how the framework I develop here can work for both explicitly moral cases, and also how it applies to practical reasoning in general. I privilege discussions of moral cases because they are the most contentious and interesting, but the points I make about them carry over *mutatis mutandis* to any other normative domain, including the prudential domain where conventions are uncontroversially taken to be commonplace and relatively unproblematic. By the time I reach Chapters 5 and 6, conventions (limited or otherwise) will have become just one example amongst others of a kind of action-guidance in which the phenomena I'm discussing can arise—in that case, the fact that individuals can do as the action-guidance recommends without appreciating why they should.

I talk about 'principles' of moral reasoning, since that is the most widespread and settled term for the family of processes and mechanisms that regulate reasoning.[5] Particularists believe

[4] Brian Skyrms, *Signals: Evolution, Learning, and Information* (Oxford: Oxford University Press).
[5] Aquinas already uses the term this way (in the Latin, 'principia'). John Finnis, *Aquinas : Moral, Political, and Legal Theory* (Oxford: Oxford University Press, 1998). 25.

that ethics isn't regulated by principles, but they use the notions of rules and standards to do similar work. While particularists, by way of their acceptance of holism about reasons, deny that rules or standards act in the same way in every case, they do accept that reasoning is informed by what the pronouncements of rules and standards in a particular case would be. Thus, they accept the importance of rules and standards that extend across cases. This is enough for my analysis—the degree to which the rules and standards are the same across individual cases is unimportant for my analysis. I ask particularists and other parties who object to 'principles' to substitute it with their favoured term for what regulates reasoning.

Limited conventions are the weakest form of convention that can hold in morality, because the existence of underdetermination is the weakest possible condition that allows any conventions at all. To show this, I'll make us of a distinction between *strong* and *weak codification*, introduced by Rosalind Hursthouse. The thesis that there is a correct set of principles that allow for no underdetermination the *strong codification thesis*. The *weak codification thesis* hold that there are true principles than can be codified in a way that applies generally, but the codification fails to avoid all underdetermination.[6] If there is any issue that isn't decided by some codification, then those remaining cases may (when strategic and repeated) become SUP cases, and would then be amenable to conventions.

Here I don't argue one way or another about whether we should accept strong codifiability. But its truth is by no means obvious, and it has been challenged for a number of decades from various corners: by particularists like Jonathan Dancy, coherentists like John McDowell, constructivists like Onora O'Neill, virtue ethicists like Hursthouse, and simply as unviable in its own right by Bernard Williams and others. This, I believe, gives me license to investigate what our moral landscape would be like if strong codifiability doesn't hold.

Lewis's analysis isn't the only one available for conventions, and while his is the

[6] Rosalind Hursthouse, *On Virtue Ethics* (Oxford: Oxford University Press, 1999). 39-42.

predominant one it has also attracted a lot of criticism. The most prominent criticisms and alternatives are from Margaret Gilbert and Ruth Millikan.[7] I don't here embark on a defence of Lewisian conventions against competing conceptions, for the simple reason that their criticisms don't threaten the manner in which I intend to make use of them. For my purposes it doesn't matter whether the appeal to Lewisian conventions offers as general an explanation of social phenomena as he wants it to be, so the counterarguments don't concern me, even if they all strike home. It isn't contested by Gilbert, Millikan, or other critics that Lewisian conventions are at least a coherent analysis and could apply to at least some social phenomena. What I do here is identify a particular kind of coordination problem, indicate what would count as a solution, and demonstrate that all these solutions are Lewisian conventions. What the standing of his analysis is outside of SUP cases simply isn't pertinent to my purposes. I do in this thesis extensively engage with the work by Gilbert and Millikan, in Chapter 6 and Chapter 2 respectively, especially concentrating on their alternatives to Lewisian conventions.

III. What is covered in this thesis

The thesis divides into broadly three parts of two chapters each. The first two chapters consist of the full presentation and defence of limited conventions in moral reasoning. The middle two chapters offer two application of the framework to familiar phenomena in our moral practices. The last two chapters investigate the effects of limited conventions on moral epistemology.

Chapter 1 is devoted to explaining the way that limited conventions supplement a community's principles. Even sincere and conscientious individuals with a full grasp on the principles that they and their fellows endorse would still need the guidance that limited conventions provide. There I provide a positive argument for why limited conventions entail moral norms to conform to them, and why not conforming to the limited convention entails

[7] Margaret Gilbert, *On Social Facts* (Princeton, NJ: Princeton University Press, 1992); Ruth Garrett Millikan, "A Difference of Some Consequence Between Conventions and Rules," *Topoi* 27, no. 1-2.

moral wrongs. I also address a number of objections to conventionalist views in morality, mainly by differentiating my view where conventions supplement principles from immodest views where the content of morality is meant to be thoroughly conventional.

Chapter 2 is devoted to explaining why some of our moral norms are partly constituted by social practices. That is, some moral norms are 'practice-dependent', to use the terms of Brennan, Eriksson, Goodin, and Southwood. These four have co-authored a view where all moral norms are practice-independent. In contrast, I develop a framework of *nested purposes*, appealing to work by Ruth Millikan and Andrei Marmor. This means that one and the same action can serve a multitude of different ends, to the effect that there is no competition between doing something because it is a convention and doing it because it conforms to a moral norm. Using this framework, I argue that the requirement that moral norms be practice-independent is misguided.

Chapter 3 presents the application of limited conventions to the justification of authority. I argue that one way (but not necessarily the only way) that commands get normative force is because they bring about limited conventions. This means that a command is a way to create the necessary expectations among those subject to the authority for them to coordinate in how they respond to an SUP case. In doing so, I present a bridge between evaluating commands one-by-one, and evaluating them as coming from an individual vested with authority. To illustrate the approach I analyse parental authority as an instance of conventional authority, despite the fact that the authority of parents isn't usually taken to be conventional. I also discuss how conventional authority gives us a way to make sense of how different overlapping authorities influence an individual (for example, how a child is subject both to their parent's authority as well as those of leaders in the wider community).

Chapter 4 moves from cases where conventions add to our moral obligations one-by-one to considering what the effect of limited conventions are on wide-ranging moral practices. I do so

by considering how conventions make sense of variation in the understanding and implementation of virtues and vices between different societies. My stalking horse there is the relativism of David Velleman, who thinks that cultural variation in recognised action-types undermines the ability of cross-cultural moral evaluations. In contrast, I indicate how we can track variations from one culture to another by way of our grasp on the evaluative point of the action-types in question. I also give an extensive treatment of how action-types relate to trait-types, since this is an ineliminable part of the everyday understanding of virtue and vice.

Chapter 5 takes seriously the fact that in the cases where there are established limited conventions, an individual can succeed at doing what is right by the lights of their endorsed principles merely by conforming to those conventions. This means that limited conventions allow for at least two different ways to learn how to do what is right: reasoning from your avowed principles and seeing that the convention plugs the gap that results from the instance it covers being an SUP case, what I call *knowing conforming*; and the far less cognitively demanding option of *merely conforming* to the expectations that constitute the convention and are available by way of common knowledge. In the chapter I give an exhaustive comparison of these different avenues, and argue that mere conforming is sufficient for conforming to your principles, even in cases where your grasp on what you should do has large gaps or even contains outright errors.

Chapter 6 is devoted to investigating the limiting case of mere conformity where an individual doesn't even have a self-understanding of themselves as working towards some particular end, even if they reliably conform to a limited convention (or some other social regularity) which secures that end. This means that individuals may reliably and repeatedly succeed at some task that they don't have a self-conception of themselves as performing. Instead, they will see themselves as conforming to that social regularity, unaware of the further purpose that it serves. I defend the possibility of this condition, which I call *rational alienation*,

give a description of how it may come about and persist. I also relate it to the work done on the origins of moral behaviour done in especially rational-choice-theoretic philosophy by Christina Bicchieri and Robert Sugden, and in the biologically-informed philosophy of Brian Skyrms and Ruth Millikan, amongst others.

IV. Who is the audience for this thesis?

i. What it offers to those working on conventions

In the first instance this thesis is a contribution to the literature on conventions. It gives a novel application of Lewisian conventions to a domain that hasn't yet been studied—to my knowledge nobody has yet tried to show how conventions can act as a supplement to background principles, nor where the purpose of conventions is to make action-guidance determinate.[8] The most similar account that I am aware of is Andrei Marmor's, who distinguishes between deep and surface conventions and indicates how an array of surface conventions can give determinate action-guidance about how to conform to deep conventions. An example of his is how in medieval Christian religious art there is the deep convention that the art is for the glorification of the divine, and then the surface conventions of how to portray saints in religious art give determinate action-guidance about at least that aspect of how to appropriately glorify the divine.[9] On his account, one kind of convention acts as a supplement to a different, background convention, whereas on my account the conventions ultimately bottom out in and inherit their import from the underlying principles. I take on board many features of Marmor's account, including this thought that there can be nested conventions in play for any one instance of cooperation; his account features prominently in Chapter 2. But despite the many overlaps between our work, my concerns aren't quite his. None of the

[8] For instance, no similar approach appears in the survey by Bruno Verbeek, "Conventions and Moral Norms: The Legacy of Lewis," *Topoi* 27, no. 1-2 (2008).
[9] Andrei Marmor, *Social Conventions: From Language to Law* (Princeton, NJ: Princeton University Press, 2009).

applications of limited conventions I survey are found in his work, and I avoid discussing his central case, conventions in the law, because that would involve engaging a different set of issues and different interlocutors than I address here.

ii. What it offers those working in metaethics

This thesis is of interest to people working in metaethics in general for at least two distinct reasons. The first is because it is commonly accepted by philosophers that there are substantial variations between the moral practices of various societies, but that these variations are laid over robust underlying similarities. It is widely accepted by philosophers that these underlying similarities reflect universal concerns that need to be addressed by any society, and the different practices are different ways in which societies address them.[10] However, the process by which this contingent development of universal concerns is meant to arise is underdeveloped in the literature, an omission this project corrects. In particular, different ways of explaining the role of conventions in societal variation will give different levels of import to the conventions, and they have different results about the extent that conventions modify moral practices.[11] Here I hope to settle at least some of these questions, at least as they pertain to variation because of the SUP.

The second reason this thesis is of interest to metaethics in general is that limited conventions indicate that metaethicists need to either establish that their favoured view is secure against SUP cases, or acknowledge that conventions are an appropriate source for action-guidance. This is the subject of the extensive treatment I give in Chapters 4 and 6 of the way individuals can relate to a piece of action-guidance that is made available to them through

[10] For instance, see the comments on this topic made in the widely-used introductory textbook by James Rachels and Stuart Rachels, *Elements of Moral Philosophy*, 8th ed. (New York, NY: McGraw-Hill). 14-31.

[11] For instance, the Rachelses downplay a large amount of societal variation as 'mere conventions' of seemingly little importance, but do give a hint as to the way different conventions can be the ways people give different expressions to the same motivations. A much thicker role for conventional variations without giving up the thought that they are different developments of the same underlying concerns can be seen in e.g. Bernard Williams, "The Truth in Relativism," *Proceedings of the Aristotelian Society* 75.

a convention rather than through their own appreciation of the principles in play.

By the same token, limited conventions provide a mechanism by which general and abstract demands can be situated in a particular community with its own contingent features and history. I here provide an analysis of how even if there is a stable ground for ethics the build-up of precedents, of some roads taken and others passed by, and simply accidents of history can lead to a community having the code of ethics that they have, and it be fine for them to do so even when it conflicts with the codes of their neighbours.

iii. What it offers those working in normative ethics

Another motivation for the project is that it is sometimes taken to be a pronounced weakness for a moral theory to leave problem cases unsettled, and sometimes philosophers take this to be sufficient reason to dismiss such theories.[12] On the other hand, the prospect that principles frequently underdetermine moral decisions has prompted considered responses since at least Aquinas,[13] continuing throughout the philosophical tradition[14] up until the present day.[15] This project falls firmly within the latter camp, against the former. I hope to describe a robust response to moral underdetermination, which goes some way to rob indeterminacy of its bite. Here I highlight the role conventions can play in serious and respectable moral systems, and show that moral systems can draw on them to give valuable guidance even in the face of underdetermination. In Chapter 1 I give an extensive survey of how limited conventions fit into existing systems of normative ethics.

---

[12] E.g. Philip Pettit, "The Consequentialist Perspective," in *Three Methods of Ethics: A Debate*, ed. Marcia W. Baron, Philip Pettit, and Michael Slote (Oxford: Blackwell), 115-17.

[13] E.g. Thomas Aquinas, "Treatise on Law," in *Summa Theologica*, I-II q. 90-97, II-II q.57-60.

[14] E.g. Adam Smith and Immanuel Kant both acknowledge and respond to the possibility of underdetermined moral reasoning (both, as it turns out, appeal to appropriately cultivated dispositions of character to guide choice in such circumstances). See Robert Shaver, "Virtues, Utility, and Rules," in *The Cambridge Companion to Adam Smith* (Cambridge: Cambridge University Press). Talbot Brewer, "Maxims and Virtues," *Philosophical Review* 111, no. 4.

[15] E.g. David Braybrooke, "No Rules without Virtues; No Virtues without Rules," *Social Theory and Practice* 17, no. 2 (1991). David B. Wong, "Pluralistic Relativism," *Midwest Studies In Philosophy* 20, no. 1 (1995). *Natural Moralities: A Defense of Pluralistic Relativism* (New York: Oxford University Press, 2006).

iv. What it offers those working in the philosophy of action

As noted above, the concluding chapters deal at length with how individuals relate to action-guidance that they are subject to by way of limited conventions, showing how this can come apart from the principles on which the conventions build and from which they gain their import. In doing so I give a thorough analysis of how such conventionally recommended actions may feature in the thoughts and deeds of people subject to them, and the distinctive considerations that are in play. In effect I am identifying a distinct domain of action, with its own grounds and consequences. Once we have shown that such a conventionally regulated domain of action is not an aberration or a mistake, we can appreciate how it may be a ubiquitous feature of our lives as socially situated individuals. For someone who is convinced of the import of conventions in moral guidance, as I am, that is what I take to be the most interesting part of this thesis.

1. Limited Conventions about Morals

In this chapter I present and defend limited conventions, a version of Lewisian conventions where the result of a convention is limited by general principles (or rules, standards, etc.), such that the conventions supplement rather than replace principles. My substantive claim here is a hypothetical: for any moral theory, if that theory faces a certain class of problem cases, any method that allows people subscribing to that theory to navigate through such problems does so by way of limited conventions. This means that conventions play an important part in our moral lives insofar as we face these kinds of problems.

The kind of problem case in question are instances of what I call the strategic underdetermination problem. I start from the observation that, unless every moral issue could be settled from general principles, underdetermination will arise, meaning there are multiple courses of action that satisfy your principles as well as any other but are mutually exclusive. In such a case it would matter which of these options you decide upon, but your principles give you no way to choose. My focus here is on the social dimension of this problem: there are situations where what you should do depends on what other people do in these cases. Keeping with the decision-theoretic literature, I'll call these strategic cases. In a strategic case, if it is uncertain what the other parties will do, then it is uncertain what you should do as well; if underdetermination is in effect, you don't know which of the available options the other parties will take; thus, given the uncertainty about their actions, you are uncertain about yours as well. That is the strategic underdetermination problem (SUP for short), and it undermines your ability to reason towards your moral ends. That is the problem for which I argue limited conventions are the unique solution.

This chapter has three sections. In the first section I introduce limited conventions by applying Lewisian conventions to the hypothetical case where principles underdetermine what

we should do. In the second section I give a range of examples of the kind of phenomena I intend limited conventions to cover, culminating in a claim that every moral theory has a need for them, insofar as it has general principles that are vulnerable to the SUP. In the third part I give a defence of the normativity of such conventions, and also respond to various objections that have been raised against similar positions.

I. Lewisian and Limited Conventions

My approach is to apply Lewisian conventions to the SUP.1 First, I'll highlight what it is that makes SUP cases of special interest.

Underdetermination is important because if for every case there was determinately a single best response, then there likely wouldn't be a need for conventions to settle which response to go for. Accordingly, the stalking horse for my work here is something Rosalind Hursthouse calls the strong codification thesis: the claim that there exists a codification of general principles such that every possible case has a determinately right answer according to those principles, with no room for doubt. She contrasts this with views that do allow for less-than-perfect determination, including weak codification, where there are true principles than can be codified, but still underdetermine some cases.2

Consider the case where there is a strongly codified decision procedure. This would be a situation where all of the parties to the repeating situation follow the same principles, follow them earnestly and conscientiously, knowing that this is true about everybody else in the situation, and those principles leave no doubt about what each party should do. For instance, simple games like tic-tac-toe are strongly codified, as would classical act utilitarianism be in the idealised condition where everybody knows exactly what the most beneficial possible action is. *Ex hypothesi*, every party playing such a game, or every agent choosing what to do on

[1] Lewis, *Convention*.
[2] Hursthouse, *On Virtue Ethics*: 39-42.

act-utilitarian grounds will be able to not only judge what they should do, in light of their avowed principles, but also what every other agent will decide on. No coordination problem would arise here.

However, if the principles are weakly codified and can be no more precise than to pick out a set of candidate courses of action with significant differences between them, then you can find yourself in the difficult scenario where there are multiple live options open to you, even granted that you are sincere and conscientious in following the principles. What is more, in strategic cases what I should do depends on which of these options you take. But for weakly codified principles I have no determinate way to predict what you will do. For example, a more complex game like chess has codified rules but they don't make it obvious in every instance what the right move is to play, meaning that I can be surprised by what move you make, sometimes to the detriment of my winning chances. Similarly, even classical act consequentialism allows that every individual can only act on what they expect the outcomes of their action will be, meaning that sometimes people will work at cross purposes and end up at a distinctly sub-optimal outcome through no fault of their own. In general, for every different way of applying the principles, there will be a different action recommended by it,[3] and accordingly a different strategy pursued in this strategic case, with the knock-on effects on the outcomes of other individuals' strategies. But the situation as described has no way for me to discover what you will do, unless there is another step in our deliberation where we are informed of each other's plans. In the absence of general principles that can give a determinate solution to every moral quandary, there is the possibility of situations where even under the best of circumstances we can't tell how the other parties are going to react, and our own reasoning is hobbled accordingly.

Lewis identifies *coordination equilibria* in his game-theoretic analysis of conventions as

[3] For simplicity's sake we are disregarding cases where different understandings lead to the same action. In these cases strategic underdetermination doesn't arise, so they aren't of interest here.

particularly important. These are the sets of strategies such that the outcome reached by them is preferred by every participant to any outcome that would result if a single participant were to change their strategy. The regularities that would be instantiated by different conventions are each a coordination equilibrium. I find it useful to also have a notion that highlights the outcomes rather than the strategies. To do so, let us introduce *benign outcomes*; these are any that isn't a *malignant outcome*. A malignant outcome is one your general principles consider determinately worse than some other available one. An outcome is available from your starting set of strategies if it could be reached by a change in strategy by any individual. One outcome is determinately worse than another if one of the principles you subscribe to says it is worse than some alternative outcome, and there is no other principle you subscribe to which says it is better than that alternative.

To put it precisely: the benign outcomes are those where there is no other outcome that can be reached if only one individual changes their strategy and the resulting outcome is preferred to the original one by any individual (the individual changing their strategy and the one who prefers the resulting outcome needn't be the same).[4] The classic example of different benign outcomes in the same situation is deciding on whether people should drive on the left or the right side of the road.

It is worth stressing that we want our ways to arrive at benign outcomes not to depend on the individuals involved explicitly telling each other what they intend to do every time. There are simply too many cases where such a conferring would be impractical or impossible. Consider what it would be like if four drivers arrived at an intersection at the same time, without a system of signalling or right-of-way to help them through it. The only way they could confer would be for each of the drivers to exit their vehicles and meeting in the middle of the

[4] So, all benign outcomes are Nash equilibria outcomes, but not vice versa. They inherit this feature from Lewisian coordination equilibria.

intersection, before climbing back in their cars and setting off in the agreed order. The less scope there is for case-by-case resolutions of instances of a SUP case, the more beneficial a regularity settling the issue would be. Common knowledge is necessary for such expectations in the absence of the parties conferring with each other each time, because it is the only source of expectations available to all of the parties across multiple instances of the recurring situation.[5]

i. Every SUP case solution is a Lewisian convention

Below is my extended argument for why any SUP case solution—any instance of a community having a regular response with which they address an SUP case—is a Lewisian convention. What is meant by 'solving' or 'addressing' an SUP case is important to note. We aren't interested merely in there being an instance of the recurring situation where the individuals involved happen to coordinate towards a mutual beneficial end. What is at stake is their doing so reliably and spontaneously. If they don't do so repeatedly, then the problem will just arise again in the next iteration of the recurring situation, and their fellows being able to depend on their playing the part is threatened to the extent that their response is unreliable. And if they don't play their part spontaneously, then it can't be said that the individuals involved have a solution to the SUP case, because they are depending on whatever external factor overcomes their lack of spontaneously coordinating. Until the individuals involved have internalised whatever prompt they are using, they still suffer from the uncertainty in SUP cases. And to avoid the uncertainty they must also have the firm expectation that their fellows won't frustrate their attempts by diverging from this set of strategies. So, if the SUP case has been addressed, individuals will both conform to the set of strategies in question and expect their fellows to the do the same. I call this a *settled response* to an SUP case. I also occasionally talk about a group

[5] C.f. Marmor, *Social Conventions*: 20.

of people showing *general conformity* to a given set of strategies or a regularity. This is just shorthand for 'almost everybody conforms, and almost everybody prefers (at least dispositionally) that almost everybody conforms'.[6] So, X is a settled response for SUP case Y is there is general conformity to X in Y.

Another important notion in play here is 'common knowledge', which Lewis has a highly developed and influential account for. Common knowledge amounts to an iterated structure of expectations to the effect that everybody expects everybody to conform to the regularity in question, they expect that everyone expects this of them too, they expect those expectations, and so on, potentially for an infinite series of ever-higher-order expectations.[7]

There are two things to note here. Firstly, and counterintuitively, 'common knowledge' needn't involve knowledge at all, but instead beliefs, and it doesn't matter whether they are true or not, just that they are generally believed across a population.[8]

Secondly, common knowledge should be read not as something that is occurrent in the mind of individuals every time they are in a situation covered by it, but instead as something available dispositionally, such that from an understanding of their place in a regularity they can refer to as many iterations of these expectations as they'd like. In almost any instance all that would be at stake is the first-order expectations that this person does one thing, that person another, etc., and in a small subset of these the second-order expectations matter, and so on for smaller and smaller subsets of cases. Higher orders of iterated expectations are available, but almost never called upon. In this respect common knowledge is like embedding in the grammar of a language: when prompted people can generate a sentence with as many levels of embedded clauses as they like, but most of the time they only attest sentences with no or one level of

[6] This is meant to mirror Lewis's phrase 'general expectation' , as in 'the general expectation of conformity to regularity R', in David Lewis, "Languages and Language," *Minnesota Studies in the Philosophy of Science* 7, no. Journal Article (1975).
[7] *Convention*: 52-60.
[8] As Lewis points out in "Truth in Fiction," *American Philosophical Quarterly* 15, no. 1: 44, n13.

embedding, and only in contrived cases more than three or four levels.

Let us move on to Lewis's definition:

> *A regularity R in the behaviour of members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P,*
>
> 1) *almost everyone conforms to R;*
> 2) *almost everyone expects everyone else to conform to R;*
> 3) *almost everyone has approximately the same preferences regarding all possible combinations of actions;*
> 4) *almost everyone prefers that everyone conform to R, on condition that at least all but one conform to R;*
> 5) *almost everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',*
>
> *where R' is some possible regularity in the behaviour of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R.*[9]

I'll now go on to show why any settled response to an SUP case meets all of these conditions in this definition. There is one proviso: I discard Lewis's requirement that the parties to the convention are aware of the different alternatives. It is easy to allow for this requirement, but I follow Tyler Burge in holding that conventions don't require its participants to know what the other possible regularities are, or even know that the regularity they are engaging in is conventional. This feature is important for my treatment later in the thesis for the epistemology of conventions and other forms of action-guidance that rely on regularities.[10]

It is straightforward to show that any settled response to an SUP case will meet conditions (1), (2), and (3) of Lewis's definition. Conditions (1) and (2) are met simply by the fact that any solution to an SUP case will be a settled response to it, meaning that the parties to the case generally conform to the same regularity. Condition (3) is met because SUP cases arise when

[9] *Convention*: 78.

[10] This is discussed in §III.iv when I address Southwood and Eriksson's 'Imelda's Inn' example, as well as in the Introduction and Chapter 2. I extensively discuss in Chapters 5 and 6 what results for limited conventions and similar social regularities if we take the weakening of this epistemic requirement seriously. It is easy to account for this condition, but I believe there is sufficient reason not to. See also the argument against this requirement in Tyler Burge, "On knowledge and convention," *Philosophical Review* 84, no. 2 (1975).

the members of a population are working from a shared set of principles that they are all sincere and conscientious in their attempts to follow. Because the individuals all accept the same principles as constraints on their actions, the preferences they have about the outcomes that result from their actions are going to be approximately the same, being the preference to conform to rather than flout their principles.

The interesting part of arguing why SUP case solutions are all Lewisian conventions is showing why they match conditions (4) and (5) of Lewis's definition, and in particular, why it will be a matter of common knowledge that everybody confirms given that everybody else does.

Firstly, a bit on the proviso about weakening Lewis's conditions, in line with the argument offered by Burge and my concerns later in the thesis. As Lewis understands these conditions, it must be that the parties to a convention have a preference over the conditional 'if everybody but one person conforms to this regularity, then that one person should conform as well'—whether this preference is dispositional or occurrent doesn't matter. The case for SUP case solutions being Lewisian conventions also goes through using this reading of the conditions, and they will cover the same range of cases that Lewis is aiming for. But Burge and I press for the weaker requirement (weaker with respect to the epistemic demands on participants) that would allow for regularities to be conventions even if individuals didn't recognise them as such, or didn't appreciate all the relevant alternatives. That drives us to read these conditions instead as the narrow-scope reading of 'if R is the actual regularity, then everybody prefers to conform to R'. That is, if we use brackets to indicate the scope of the preferences in question, the weakened version of the condition reads:

> Weak Reading: *if it is the case that R is the actual regularity, then everybody prefers (general conformity to R)*

and not:

> Strong Reading: *everybody prefers (that if it is the case that R is the actual regularity*

then everybody prefers general conformity to R)

This covers condition (4). The same goes *mutatis mutandis* for the counterfactual preference for general conformity to R` if it was the actual regularity instead, corresponding to condition (5).

I offer the following argument for why any SUP case solution will meet Lewis's conditions (4) and (5):

A. *The parties to the SUP case solution need to have some other shared resource that leads all of them to the same benign outcome.*
B. *Any such resource entails the expectations of general conformity to the regularity.*
C. *Any such set of expectations about general conformity to a settled response would count as common knowledge.*
D. *Thus, any settled response to an SUP case will be a matter of common knowledge—satisfying Lewis's conditions (4) and (5).*

Since conditions (1), (2), and (3) have also been met, we can conclude that every SUP solution is also a Lewisian convention.

Let me argue briefly for the truth of the premises. Premise A follows from how the parties to an SUP case would need to agree on how to apply the principles, but cannot depend on the principles to guide a unanimous choice. Thus something else will also be needed.

Premise B introduces the important move that we don't necessarily try to directly use conventions to solve SUP cases. Instead, we adopt an agnosticism about how it is that the parties come to coordinate, and make the observation that as people come to depend on this coordination, they will come to have a corresponding structure of expectations. This is because as they expect people to draw on this resource, they also consequently expect people to do as this resource recommends, which is conform to the regularity.

Premise C is a bridging premise, to go from Premise B which uses my term 'settled response' to Lewis's definition which doesn't use that term. It follows straightforwardly from my definition of a settled response that any settled response will be the object of common knowledge. As I discussed

above, since there is general conformity to a regularity, there is the general expectation of conformity. The rest of the iterated expectations that make up common knowledge follows by the same token. So, we arrive at our conclusion at D that all SUP case solutions meet Lewis's conditions (4) and (5), which together with their meeting (1), (2), and (3) entails that all SUP case solutions are Lewisian conventions.

ii. Moral reasoning from conventions

The notion of preference which Lewis uses requires only that if someone prefers doing x to y, *ceteris paribus* they would do x instead of y.[11] This is thin enough to allow me to recast subscribing to a moral principle as a preference to act in accordance with that principles rather than contrary to it. This isn't to turn moral preferences into prudential ones; it is to respect the fact that *ceteris paribus* you should do what is moral rather than immoral. Being a sincere and conscientious moral agent involves doing that at the very least. Using the same sensitivity to strategic considerations as in prudential cases, we can identify strategies that lead to benign outcomes depending on how they conform to our principles.

Limited conventions should be seen as the product of a two-step procedure: first, the benign outcomes are identified, and second, one of them comes to be the conventional outcome (through an intentional process or not). To use terms introduced by HLA Hart, the first step is *content-dependent*, where the selection is made with reference to standards of evaluation independent of the decision-making process, and the second step is *content-neutral*, where the decision-making process itself determines the correctness of the choice. Conventional theories are paradigmatically content-neutral, because an outcome is supposed to be the correct goal simply because it is the one recommended by the convention in question. But in my analysis this is just one part of the story. This chapter introduces how conventions could also be content-dependent, by way of how general principles can structure the options that we select over in a

[11] David Lewis, *Convention: A Philosophical Study* (Malden, MA: Harvard University Press, 1969). 9, 90.

content-neutral fashion.

An example of this is the recently implemented requirement in the United States for individuals to be covered by health insurance. This requirement can be very burdensome, in for instance the bureaucratic workload both of the government and of individuals who are required to comply, and it is easy to find fault with the system. And other ways to secure something approaching universal access to effective health care are available.[12] But these mandated forms of insurance serve a genuine purpose. The motivations for implementing this system has often been explicitly moral, with many references of the duty of the state and of citizens to make effective healthcare achievable for their fellows—something whose status as a general principle is I think not in any doubt. That principle has ruled out the many options, including the old system which has by very common assent been judged as not sufficient for the purpose.[13] This is a content-dependent step. It has left a range of options that are at least minimally sufficient, and then one of them was selected. My claim is that it doesn't matter how that choice was made, because now in the US there is a system of (near) universal healthcare which is a going concern and is in actuality how the US discharges its duty of healthcare to its citizens. Accordingly, all those who fall under the US's jurisdiction have a duty to comply with that system, since individuals cannot by themselves implement, say, a single-payer system, but through their non-compliance they can undermine (to some small but real extent) the ability of everyone to reach their morally required end. This happens without reference to how the choice was made, and thus is a content-neutral step. What is more, on my analysis this requirement is

[12] Even in the US itself, where a certain class of the population qualifies for a scheme very much like the single-payer schemes prevalent in other wealthy nations: enrolment into a no-fault insurance scheme with automatic enrolment, run with the US government as a single payer.

[13] Healthcare costs for most of the population being a responsibility of the individuals involved, insurance being left to their discretion, and emergency services not refusing service to people who can't pay, but still charging for this service. This is a very bad system. One reason is because it is literally ruinously expensive to people without insurance, and to a lesser extent to the providers of emergency services. Another is because there is a lot of essential medical care which isn't done through emergency services, like preventative medicine and treating chronic conditions.

held in place by the expectation that it is by way of mandated insurance that the US complies with the principled requirement to provide healthcare, meaning that notions like sanctions against non-compliers or the special authority of the law are superfluous to the normative standing of this case, whatever their virtues may be. This is an instance of where we can find good independent reasons to comply with the law.

II. Examples of limited conventions

Let me offer a few putative examples of things discussed by moral philosophers that I take to be instances of limited conventions. First, I'll identify some existing meta-ethical positions that are amenable to such an analysis, then I'll sketch out how limited conventions can play their part in a number of contemporary ethical theories.

i. Limited conventions in the philosophical literature

My analysis of limited conventions is one out of a long tradition of attempts to find a place for conventions in moral reasoning. But we should immediately note the most important point of difference between this approach and most attempts to find a role for conventions in morality. Conventionalist theories normally try to describe how conventions arise independently of whatever moral principles are in effect—a fact that Bruno Verbeek uses to question the suitability of conventions for determining moral norms.[14] Limited conventionalism isn't one of these theories. Here I take the existence of some set of principles as a given, identify a problem when reasoning from those principles, and use conventions to offer a solution.

I present four examples of positions within the literature which I cast as examples of limited conventionalism. The oldest such model I have found is the reconciliation of human and

[14] Verbeek, "Conventions and Moral Norms," 82-85.

natural law in Aquinas.[15] He offers a theory of what he calls a *determinatio*: selecting one amongst a range of specifications allowed by abstract, underdetermined principles in order to make the principles practicable. Aquinas's central example is of someone designing a house: a house needs a floor, a roof, walls, doors, windows, and so on, but the various requirements of a house don't determine any particular design. In order to build a house, you need a particular and determinate design from within the range that satisfies the basic requirements.[16] Aquinas discusses strategic choice alongside non-strategic choice without a strong distinction between these (the choice of which house design to build is generally non-strategic), but strategic considerations are cited as a reason to make use of *determinationes*, and for these cases his analysis is largely in step with mine.[17] Pointedly, *determinationes* can be understood as the product of a two-step selection process, where the first is finding the range of options that are sufficiently consistent with the general principles, and the second is selecting arbitrarily among the limited outcomes (in Aquinas, usually through a pronouncement by some authority).

A contemporary and more detailed example is David Wong's pluralistic relativism.[18] Wong claims that while there is a non-relative foundation for ethics—a series of necessities that any system of ethics would need to provide to its adherents—this radically underdetermines the form and content of our morality. To actually instantiate a social system which meets those necessities, we would need to make use of a suitable specification, since without it the members of a society need to match their fellows in a particular strategic arrangement in order to meet the basic needs, but not knowing which of the multiple available such arrangements is in force. Wong argues that this is sufficient to establish a form of moral relativism, since where the specifications of two societies differ from each other there are different moral obligations in

[15] It is possible that an older example is work in medieval Islamic philosophy about how different traditions of revelation may lead to differences in the moral demands placed on the members of the societies in question, by e.g. Al-Farabi.
[16] Aquinas, "Treatise on Law."
[17] Ibid.
[18] Wong, *Natural Moralities*; "Pluralistic Relativism."

each of them. What Wong has proposed is that an example of a society's social system is the product of a limited convention. The specific arrangement of any society is an SUP solution that is the product of a two-step selection, where the first step of the selection allows only moral codes which fulfil the basic functions, and the second makes it arbitrary which of these codes is in effect in any society. And which arrangement is in play is obviously common knowledge, with each of the adherents raised in such a system and knowing that their fellows have been as well.

Another, very similar, example is the society-centred theory of David Copp. Copp provides a theory where there is a single overarching framework to morality, which he identifies as cross-societal needs that moral codes address. But the particular features of a society make a difference to what moral code this framework recommends. In this way Copp wishes to have a variety of moral realism which embraces rather than diminishes the role of cultural variation. The particularities of a society—in the first instance its society-specific desires, but also its history, material conditions, relations to other societies, environment, and so on—are inputs in the single overarching framework, and lead to different moral codes commensurate with the differences in their situation. This means that Copp sees the development of a moral code as the product of a limited convention: the cross-societal needs constraining the range of justifiable moral codes is the first step of the selection, and the society-specific desires, including ones that are arbitrary in the sense of not being derived from determinate principles, further narrow down the options to the point that a specific moral code becomes selected. Copp at one stage defended the view that for any one society there is a single moral code that is justified—the one that fits best with the cross-societal needs and society-specific desires. [19] Copp has since given up on this requirement, and allowed that it is possible that there are multiple codes that could be

[19] David Copp, *Morality, Normativity, and Society* (Oxford: Oxford University Press, 1995). Normativity

justified within a given society.[20] Like the later Copp, I believe that we shouldn't expect that there will be a unique determinately correct moral code for a society—and discussing SUP cases is discussing one example of how this determination isn't guaranteed. In my terms, the benign outcomes in an SUP case are products of what Copp calls the 'best codes' among which a society chooses.

The two examples above are views of how morality as a whole can be a product of a constrained arbitrary selection. But this isn't what I take to be the central example of limited conventions. What I think is more important and prevalent is when there are general principles that need not be derived from convention, but which you require conventions to put into practice. An example of this is David Braybrooke's argument for why act-utilitarianism collapses into rule-utilitarianism in anything except the most idealised conditions.[21]

Braybrooke points out that there are going to be occasions when agents, through lack of time, information or foresight, have to make choices without making full use of act-utilitarianism as a decision procedure. When the agents meet with their fellows in more favourable moments, what Braybrooke calls the 'community-in-session', they can evaluate how people acted in these sub-optimal conditions. Whatever the result of that evaluation, the precedent the community-in-session sets will produce a rule: if the past action is commended, then the rule is to do the same if you come into such a situation; if it is judged to have been wrong, then the rule is to do otherwise. In this way, Braybrooke argues, act-utilitarianism becomes rule-utilitarianism in the face of non-ideal conditions.

What is of interest to us here is how Braybrooke's rules are examples of a structure of expectations determining for individuals how they should respond to uncertain cases. The

[20] *Morality in a Natural World: Selected Essays in Metaethics* (Cambridge: Cambridge University Press, 2007). 243.
[21] David Braybrooke, *Utilitarianism: Restorations, Repairs, Renovations* (Toronto: University of Toronto Press). 14-19.

breakdowns of the idealising conditions mirror breakdowns in strong codifiability, where knowing the relevant principles fails to settle what you should do. When the cases in question are strategic, which they won't always be but will be some of the time, then these are SUP cases. Again, we can model the process in question as a two-step selection process, but in Braybrooke's presentation the epistemic order is the reverse of the determinative order I have presented.[22] On Braybrooke's (epistemic) way of counting, first comes the particular judgement of an individual about how to handle a particular situation, then second comes to pronouncement of the community-in-session on whether in recurring cases to follow the particular judgement, or instead some other one. On the (determinative) way of counting I have presented here, the first step is seeing whether the agent's choice is consistent with the general principles (in this case, the one-member set of principles of act-utilitarianism), and the second is the agent's decision that selected that particular option from among those in sub-optimal conditions, a choice that forms a precedent for others to follow in the same situation (or to reject in favour of some other option that then becomes regular behaviour). The pronouncements of the community-in-session cement these rules as common knowledge. Thus, these rules are limited conventions.

ii. Limited conventions supplementing Kantian ethics

Kant isn't typically the theorist you look to for models of contingent and socially-situated moral theories.[23] But there has in recent decades been a move towards this aspect of his moral thought, especially by theorists who highlight the work in his *Metaphysics of Morals*.[24] But

[22] More on the epistemic and determinative orders involved in conventions in Chapter 2, section II.ii.

[23] The lack of development of his theory in this regard is the grounds for Hegel's criticism of his ethics. Robert B. Pippin, *Hegel's Practical Philosophy: Rational Agency as Ethical Life* (Cambridge: Cambridge University Press, 2008); Robert Stern, "On Hegel's Critique of Kant's Ethics: Beyond the Empty Formalism Objection," in *Hegel's Philosophy of Right: Essays on Ethics, Politics, and Law*, ed. Thom Brooks (Malden, MA: Blackwell, 2012).

[24] See Barbara Herman, *Moral Literacy* (Cambridge, MA: Harvard University Press); Onora O'Neill, *Towards Justice and Virtue: A Constructive Account of Practical Reasoning* (Cambridge: Cambridge University Press, 1996).

even in the *Groundwork for the Metaphysics of Morals* we find something related to the strong/weak codification distinction at work in his distinction between perfect and imperfect duties, which similarly is a distinction based on the definiteness of the guidance involved.[25] The imperfect duties are those that aren't uniquely determined to be required by the categorical imperative, and accordingly aren't strongly codified either. So, since there is underdetermination here, there is scope for SUP problems. And they are easy to find. While Kant's central example of an imperfect duty, the duty of charity, isn't strategic in many of its instances, there are cases where the duty of charity needs to be discharged but your ability to do so depends on co-ordination.

Consider providing wedding gifts for a young couple. Since a marriage is the occasion to celebrate the founding of a new household, and because of the help most new young couples need to furnish their household, such furnishings are standardly and properly included as wedding gifts. But any household only needs a limited number of its various furnishings, so a couple who gets given five bread-makers has at least four too many. Accordingly, the convention of wedding gift registries has arisen, to coordinate what gifts are given. Nobody can be said to be under a perfect duty to either give or not give a bread-maker, but nonetheless for a particular wedding there are circumstances where it would be determinately wrong for you to give a bread-maker—if there is a gift registry and the bread-maker has already been marked off. This is an SUP case, because there are many different gifts a young couple can appropriately be given and even more mappings of gifts to givers that would be appropriate (the Smiths give a bread-maker, the Thambos a set of cookware, and the Ul-Haqs fine linen; or the Smiths the linen, the Thambos a bread-maker and the Ul-Haqs the cookware; and so on). So, there is an imperfect duty of charity to give something appropriate to the young couple, and in

[25] Similar, but not identical. Codification relates to the content of principles, whereas the categorical imperative on which Kant bases perfect/imperfect duties is purely formal.

discharging it reference needs to be made to the limited convention implemented by a wedding registry.[26]

iii. Limited conventions supplementing virtue ethics

Among the views surveyed here, virtue ethics is likely to find the possibility of SUP cases the least surprising and threatening. There is no general expectation of strong codifiability amongst virtue ethicists, and they are often the theorists who have pressed the limitations of codifiability in the literature.[27] But here is an example of how we don't just have a breakdown of codifiability, but also a population-wide regularity in behaviour that leads to a coordinated and conventionally-established response to the issue in question.

Take Aristotle's now-unfashionable virtue of magnanimity (*megalopsuchia*). Many of us who aren't inhabitants of Aristotle's ancient Athens find unattractive the thought that we should cultivate the disposition to give lavish gifts or volunteer for demanding services and also to demand recognition for them. This is especially so given that for Aristotle this explicitly links with the domains of things it is appropriate for someone to get angered by: if I give a lavish gift and you pooh-pooh it, it would by Aristotle's lights be a mistake for me not to get angry at you for this slight. This is not what most of our contemporaries recognise as virtuous behaviour. But there is a point to Aristotelean magnanimity: in Aristotle's Athens there was no universal taxation and accordingly by our standards there was very little public expenditure; instead large public works or public ventures where paid for by private individuals making gifts to the city,

[26] Kant has an argument on these lines in the *Doctrine on Right* about how the institution of private property becomes enacted in a society, but it would require more space than I am willing to give to discuss it. Kant considers this to also be derived from the perfect duty to conform to the commands of your sovereign, and I would need to discuss at length how to separate the perfect and imperfect duties involved there. But I note this as an example of interest in his work. Immanuel Kant, *The Metaphysics of Morals*, ed. Mary J. Gregor (Cambridge: Cambridge University Press, 1996).

[27] Hence the aptness of my citing Hursthouse to draw the strong/weak codification distinction. See also, Julia Annas, *Intelligent Virtue* (Oxford: Oxford University Press, 2011); Daniel C. Russell, *Practical Intelligence and the Virtues* (Oxford: Oxford University Press, 2009). For this same reason many virtue ethicists also endorse particularism, e.g. Christine Swanton, *Virtue Ethics: A Pluralistic View* (Oxford: Clarendon Press, 2003).

and these ventures were led by volunteers from among the higher social classes. This covered everything from the construction of the Parthenon to military expeditions like those in the Peloponnesian War. What Aristotle is doing when describing the virtue of magnanimity as he sees it is the disposition required by individuals of high social standing to play their part in the political life of the city.

As social circumstances change, so too do the ways and means of political action, and the appropriate dispositions for public figures to cultivate. These days we have large schemes of universal taxation to finance public expenditure, so lavish gifts play no large part, and we have bodies of professionals who occupy the roles required by government. This is less magnificent, perhaps, but more reliable, and the kind of large-scale public programmes that is a hallmark of our society would be impossible without it. Accordingly, while we may still have some attenuated version of magnanimity in our evaluative vocabulary (applied to being a patron of the arts, say, or engaging in charity efforts), it is diminished in importance from Aristotle's.

Limited conventions enter the picture because the usual means of addressing the needs of public bodies is a strategic matter: if the political system depends on a universal system of taxation, then the virtues of public life will demand that; if the political system depends on gifts from private individuals, then the virtues of public life will encourage those. In a system that is in this respect between our world and Aristotle's, say, the Napoleonic world of universal taxation but where public office was held and maintained by the members of the higher social classes, the virtue of public life becomes 'public-spiritedness', and is made reference to throughout the period. The background principle of individuals needing workable public bodies stays constant, but the way that principle becomes applied differs. So, the provision for public ventures is an SUP case, and insofar as the cultivation of personal dispositions plays a part in it, the corresponding virtue of public life is a limited convention. It is important to note that insofar as personal virtues are allowed for in a theory (and they are in almost every theory) this

kind of story isn't restricted to virtue ethicists, though it has a special relevance for that theory.[28]

iv. Limited conventions and contractarianism

Here is an example for contractarianism. David Gauthier in his *Morals by Agreement* proposes a contractarian standard to handle issues about the distribution of resources that voluntary exchanges by way of an open market aren't suited to cover, to do with the relative costs to the individuals involved.[29] Gauthier meant this to handle any and all problem cases, but Robert Sugden has demonstrated that it won't do. Sugden has shown how to construct problem cases where there are multiple possible equilibria that are equally attractive according to Gauthier's distributive standard, and therefore that standard isn't able to decide between these equilibria.[30] Thus, just as market mechanisms didn't succeed in strongly codifying all distributions of resources, so too did Gauthier's standard fail to strongly codify all those cases that are left.

Since Gauthier is explicitly dealing with a domain that is ineliminably strategic, the class of cases Sugden has identified are all SUP cases, where the different benign outcomes are the different equilibria available. And as discussed, every time they have some settled way to handle a recurring SUP case, that will be a limited convention.

The Gauthier-Sugden example can also serve as an example of how we can have an aggregation of multiple layers of limited conventions. One way this could go is that Gauthier's standard might get adopted as a limited convention to cope with breakdowns in market mechanisms for distributing resources. Then, after Sugden's cases have entered the picture and attracted settled responses, we would have limited conventions to handle those in turn. So, we have the general class of resource distribution problems which in the first instance Gauthier

[28] Chapter 4 is devoted to this kind of conventional determination of the content of virtues and vices.
[29] David P. Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986).
[30] Robert Sugden, "Contractarianism and Norms," *Ethics* 100, no. 4.

wants to use market mechanisms for, then for a subset of these the limited convention of his distributive standard, then for a narrower subset of those the limited conventions for handling Sugden-type cases.

v. Limited conventions supplementing consequentialism

Finally, let us consider consequentialists and limited conventions. The Braybrooke example surveyed in §II.ii is one indication of how this may go, but the possibility is more general than that. For one thing, there is no reason a restrictive consequentialism such as Pettit and Brennan's can't make free use of limited conventions, where the principle in question is whatever the consequentialist metrics are that the theorist accepts and the convention is the plan for some set of instances the restricted consequentialist has adopted.[31] But the call for limited conventions in consequentialism is still more general than that. Even for an unreconstructed act consequentialist there is likely to be a use for them. It is one thing to say that in principle there is a determinative ranking of all available options that settles what should be done; it's another thing entirely to insist that this ranking is available in perfect detail to every individual in every instance. Patently it isn't, and consequentialists aren't shy to admit this or make allowances for it. Let us take a very mild, possibly the mildest, weakening of the epistemic position act-utilitarians could be in: that they have good information about the consequential upshot of all their available actions, but only within a set margin of error. This means that where ideally they would be able to determine the single best action (or range of equivalently good actions), they instead can only identify a range of best candidates. What is more, their estimation of what other act-consequentialists will do is subject to the same uncertainty, so they can't depend on what options others will pursue because (concomitant with the margin of error) those others may have different options that appear more beneficial to them. So, coordination among act-

[31] C.f. Philip Pettit and Geoffrey Brennan, "Restrictive consequentialism," *Australasian Journal of Philosophy* 64, no. 4 (1986); Pettit, "The Consequentialist Perspective.".

consequentialists would then be more problematic. In strategic cases, this leads to an SUP problem. So, act-consequentialists faced by this kind of margin of error will find that they can best promote their desired consequences by arranging for limited conventions to handle such cases if they re-occur, and to conform to an established limited convention in all cases where conforming to it would be more beneficial than trying to re-arrange it anew, which is likely to be very often. And as for the strictest kind of consequentialist hampered in the mildest way in discharging the demands of their theory, so for every other kind of consequentialist and any other kind of hampering as well.

III. The Normative Force of Limited Conventions

I want to reiterate Lewis's point that conventions of the type he analysed are a species of norm, that is, that to the parties involved they count as reasons to do what the convention recommends and not do anything else.[32] The wrong in not doing what a conventional solution to a SUP case recommends is that you are failing your fellow agents when they depend on you to act one way rather than another. This would undermine the grounds on which your fellows reason towards their preferred outcomes, where the relevant preferences include acting in accordance with their avowed moral principles. I argue that your fellows have every right to expect you to act in the conventional manner, and you have no good reason not to. Accordingly, for you to fail to do what is expected would be to willy-nilly frustrate your fellows.

There may be a worry that whether you find a particular limited convention compelling depends on whether you are sufficiently attracted to the good that is secured by the cooperation that results from the convention in question. Perhaps in some particular case you don't think doing what you're expected to has much to offer you, so you can question whether you should participate in it or not. But this kind of noncommittal attitude to limited conventions won't do. If

[32] Lewis, *Convention*: 97-100..

there is a question for an individual whether to conform to the convention, there is question for everybody else about whether that individual will do as depended upon, and this would undermine the expectations necessary for the convention. If this were the case, the problem that a convention solves has come back, only at a different level: we again have uncertainty about how the other people will act, but now not directly because we don't know which of the available options the other parties will take, but because we don't know whether they will or will not follow the expectations that have as their object the underdetermined range of options. This means that if there is to be any cooperation at all in SUP cases, it must be through a limited convention (for the reasons surveyed in §I.i).

While the benefits to any one individual at any one time might be modest, they are still significant. It might not be the case that every party gets what would be the best possible result for them bar none, nor that these conventions won't weigh heavier on one party than another. Nonetheless, no course of action can be picked out wherein some party would follow their principles better by not following the convention. There is a balancing of preferences in such a way that no party is injured and all are benefited by the convention, even in the minimal case where the only injury that is avoided is the uncertainty of underdetermination and the only benefit the avoidance of them. But this uncertainty is a real problem, and a solution to it a real benefit. The fact that conventions offer only a piecemeal solution to uncertainty, dealing with it in a restricted range of cases but not in others, is no objection. You don't tear up your coat because it's enough to keep you warm only in the valleys but not up in the mountains. Similarly, you don't abandon something that helps you only in some cases in favour of receiving no benefit in any case. The parties to a convention receive a real benefit in the face of a real problem, and that is enough to cement their importance.

That is my account of the normativity of limited conventions; now on to responding to objections.

iii. Conventions offer more than just a salient option

People sometimes press the objection that even if a convention offers an especially salient way to address a coordination problem, that doesn't entail that it is obligatory to follow the convention, There may be other possible conventions which one of the individuals in question may prefer, or there may be something distasteful to them about this convention. An especially sensitive treatment of this objection is given by Mark Murphy when discussing the authority of the law.[33] He doesn't make the point in terms of conventions, but conventions are an example of an arrangement that is meant to gain its normative standing from its coordinative function, so his argument applies to my view as well. Similarly, I believe that the following response works for Lewisian conventions in general, but the response is especially effective for limited conventions.

Murphy presses the point by giving an argument that cedes a lot to the coordination theorist, and purports to show that even after these concessions the fact that there is a salient way for individuals to coordinate towards a mutually beneficial end isn't enough to make it obligatory for them to conform to it. Murphy allows that there are many issues of great importance where coordinating with your fellows is a sufficiently, even maximally, good way to handle them. He also allows that for many important examples of these issues a failure to coordinate will lead to bad, even disastrous, results. But, Murphy notes, this only makes it especially pressing that the people involved come to some kind of coordinative arrangement, not that they must follow the one in question. He highlights this gap by way of contrasting how many theorists will accept that the law should be authoritative because it is a salient coordinator but not accept that husbands should be authoritative in a household as a salient coordinator. This is even though in both cases there are many benefits to be gained this way, many harms to

[33] His target is versions of the natural law theory which emphasises the coordinative function of codes of law. Mark C. Murphy, *Natural Law in Jurisprudence and Politics* (Cambridge: Cambridge University Press, 2006). 102-11.

be avoided, and accepting the husband as the head of the household is a perfectly fine coordinative arrangement, made salient by its standing as the default throughout much of the histories of the parties to the debate. But that isn't enough to make it the case that accepting the husband as the head of the household is required, since there are familiar reasons not to. This means that salience as a coordinative arrangement isn't enough to secure normativity.

My response is to highlight is to cede the point that salience isn't enough, but to stress in response that conventions aren't just salient coordinators. Conventions aren't just possible ways a group can handle an issue; a convention is something that is in place when people already have a settled response to an issue. They are going concerns, not prospective but actual. This means that not conforming to the convention isn't a neutral option. The choice that an individual needs to make isn't whether to participate in the on-going convention or make use of some other coordinative effort—the choice is whether to conform to the on-going convention or to frustrate it. It's only by way of conforming that the individual has the prospect of securing the end that the convention coordinates towards.[34]

Wouldn't this then mean that we should insist that wives take their husbands to be the head of the household, just as earlier I insisted that those in the US participate in their mandated insurance scheme for healthcare? It doesn't, for a number of reasons. Most important is that it is appropriate for every household to independently form the structure of authority within it, whereas it is impossible to settle how to finance healthcare every time someone makes use of it. In cases like the funding of healthcare what is at issue just is how to handle these cases in aggregate. But the interest in arranging a household doesn't extent to the aggregated case in this way. At the very most, wider society has an interest in households having some kind of structure for the making of decisions, without any reference to which individual occupies a

[34] This isn't to say that free-riding is impossible, because Lewis explicitly makes allowance for less-than-perfect compliance. Rather, it is to say that gaining the end coordinated to by a convention without conforming to it is free-riding, and not a different way to secure the ends that others secure by conforming.

decision-making role, but even this is hard to motivate. Wider society certainly has no compelling interest to cement the husband's supremacy over a wife. Trying to have a rule for the arrangement of households is one prominent way in which the harms of making women subservient to men arises. So, the kinds of interests at stake in the healthcare case and other instances of proper coordination aren't at stake in the arrangement of households, and there are good reasons not to adopt this kind of widespread coordination in those domains.

The same point goes against the thought that conventions are unlikely to be binding because individuals may find the current convention distasteful or would prefer a different one. An individual's personal attachment to one or another coordinative arrangement can accomplish nothing on its own—that just is what it means for a situation to be strategic. To refuse to participate in an established convention because you would prefer a different one is to choose failure over success—you frustrate the convention by not participating, and the convention is the only available means to accomplish the end of coordination. In this respect limited conventions are especially well-placed to answer the objection: the end in question is to conform to your principles, which you have antecedent reasons to do and antecedent reasons not to do something different. Insofar as the principles in question are binding, so too is the limited convention that allows you to conform to them. This point is repeated in my response to the next objection—it is a vital part of my case for limited conventions.

It must also be said that the import that many commentators place on salience in Lewis's analysis is misplaced. Lewis introduced salience as a suggestive possibility of how people can come to surprisingly high degrees of coordination without availing themselves of explicit agreements, and as a link to relevant work done in a different field. This happened against the backdrop where some philosophers (including Lewis's doctoral supervisor, Quine) seriously defended the thesis that the conventionality of language is impossible because it would require agreements, and agreements in turn require some antecedent language already being

established. Salience serves its role admirably as a suggestion about how coordination can happen without agreements, but it is by no means a pivotal part of the theory. It isn't mentioned anywhere in the definition of convention, nor is it one of the kind of coordinative arrangements that Lewis compares and contrasts conventions with. Certainly it isn't the case that common knowledge is meant to guide people to a particular strategy because it makes that strategy salient. It does make that strategy salient, but only because that strategy is the one that is the subject of the expectations that constitute common knowledge, and those expectations are what guide you, which they do perfectly sufficiently on their own without any reference to salience.

iv. Limited conventions don't arise too easily

On my account, the normativity of limited conventions comes from a mixture of the parties to a situation being expected to act a particular way, and their all acting as expected leads to an outcome that is valuable by the lights of their shared principles. It may be objected that this means that obligations may arise too easily, where every minimally desirable (by the lights of the principles) outcome would become binding in case the expectation arose that people will act in a way that would produce that outcome. This would make the view excessively demanding, or at least very implausible, by dint of the sheer number of obligations that it generates. Someone making this objection may be happy to admit that there are very important cases where it is plausible that obligations specified by limited conventions are binding, but worry that many minor obligations will turn out to be just as binding. At best this would, as it were, clog up the normative space with a mass of obligations the burdens of which outstrip their import; at worst, this would strip us of individual freedom to an enormous extent.

An example of what the objector may have in mind would be the fact that Kant had a habit of going on a walk which he kept to such regularity that his neighbours used it to set their clocks. While Kant going on his walk isn't a strategic situation, the arrangement where Kant goes on his walk and this is the signal for his neighbours to set their watches is strategic. The

worry is that Kant may be obligated to have a walk at that time since his neighbours expect him to do so and the resulting regularity allows them to set their watches, and Kant's conscientiousness would oblige him to conform to this regularity. And while Kant was famously conscientious, it seems wrong to think that the fact that his walks were that regular in a way that was beneficial to his neighbours made him obligated to continue as he did before.[35]

There are a number of things to note in response to this worry. Most importantly, it is not the fact that someone is expected to act in a particular way which forms the obligation, but instead the obligation comes from a principle in the background. This means that, since the convention inherits its normative force from the principle, it is only going to produce obligations that the principles demand in their own right. Since individuals are already committed to the principles, it isn't that the convention makes obligations out of thin air, but instead makes it possible to accomplish things which individuals were already committed to. Since there is no antecedent obligation for Kant to arrange his daily schedule in ways convenient for his neighbours, there is no consequent obligation for him to continue going for his regular walk.

In response, there may be the complaint that the obligations in place without the aid of the convention may be fine, but the extra ones that result from the conventions are not. But if the obligations that arise from a principle are in some respect odious, then it is the principle which is to blame, not the convention, since the convention can only arise for things that the principle asks for in its own right. On this score, consider the large literature of thought experiments where we are asked to imagine situations where we make trouble for some candidate principle by showing how in that situation it would require us to do things we think people shouldn't be required to do, e.g. how consequentialism may licence discrimination against a minority if they are small enough and the advantages gained by the majority in discriminating against them is

[35] I return to this example in Chapter 2, §I.iii.

large enough. The conclusion here is meant to be that we shouldn't adopt consequentialism; not that we should stop such minorities from arising. Similarly, if a convention seems to show that a principle leads to an odious outcome, then we should blame the principle, and not the circumstances which make that consequence of the principle possible. A limited convention is a means to conform to your principles, no more and no less, and this is enough to make them binding, since in SUP cases no other means is available.[36]

v. Conventions aren't estoppel

As indicated above, while conventions involve certain expectations, it is a mistake to think that the existence of those expectations is enough to establish a convention. One way this mistake can be manifested is by considering conventions as a kind of estoppel such is common in the law: where you are obligated to do something because you have caused others to have the obligation that you would do so, for instance, not going back on a promise you've made earlier. Joseph Raz has objected to making use of estoppel in order to explain the normativity of conventions as unconvincing. Raz objects that in general there is no significant way in which the parties of the convention cause each other to have these expectations. It is true that the expectations exist because the convention exists, but for almost every convention the person who caused the convention to exist is not party to the situation. Accordingly, in many situations nobody present caused any of their fellow parties in a convention to believe that they will act in the conventional manner, and thus the harm done by not following convention is not the same as falling afoul of estoppel. Nor, for that matter, is it like breaking a promise.[37]

I grant that in a convention the expectations your fellows have about how you will act are not caused by you, at least not necessarily. And consequently I grant that the harm of breaking a convention is not the same as the harm of breaking a promise. But conventions are to be

[36] Also relevant here is the discussion of how limited conventions give exclusionary reasons in Chapter 3, §III.i.
[37] Joseph Raz, *The Authority of Law*, 2nd ed. (Oxford: Oxford University Press). 237-41.

respected nonetheless. The reasons for this are very direct: as has been noted above, the existence of these conventions allows the community to coordinate towards valuable ends, allows individuals to reason towards what they value through what would otherwise be an impassable mire of uncertainty, and does so while respecting everything they value in virtue of their commitment to the shared moral standards. These expectations are to be respected because the uncertainty which their absence causes harms everybody, and we should not harm our fellows. In other words, conforming to the expectations are the means to securing ends, not the end itself. This is distinct from, and weightier than, the requirement not to frustrate the expectations you foster amongst other people, as in estoppel.

vi. Conventions aren't just products of individuals' interests

Another objection of Raz's is that conventionalist accounts can perhaps balance individuals' subjective preferences, but cannot handle social cooperation towards shared goals. It might be that the shared goal is not to the benefit of most or even any of the parties in a co-operative effort: Raz gives the examples of a community banding together to give aid to foreign victims of a flood, or for the protection of an endangered species. But without each individual drawing a tangible benefit from the coordination, depending on subjective preferences to justify conventions would leave such examples of cooperation mysterious.[38]

This is an objection limited conventionalism is especially well-equipped to answer. The relevant type of preference for limited conventions is to comply with existing moral demands. The benign outcomes that are chosen between are the outcomes that cohere best with the general principles accepted by the members of the community, crucially including those principles affecting not only members of the community. Raz's objection takes the existence of such principles as a given—the objection is that conventions make acting from such principles

[38] "Introduction," in *Authority*, ed. Joseph Raz (New York, NY: New York University Press), 6-10.

mysterious—but if a community endorses such principles then limited conventions will account for them. The applications of non-members-regarding principles are just as vulnerable to uncertainty as any other. This means that these non-members-regarding principles can be supplemented by limited conventions in just the same way that any other principles are. This means that the goods aimed at by the non-members-regarding principles are exactly those that Raz is trying to contrast to the narrow domain of subjective preferences, but are also secured by limited conventions. So, there isn't the kind of split between limited conventions' ability to handle members-regarding and non-members-regarding principles that the objection depends upon.

vi. Responding to Southwood et al's objections

Nic Southwood is someone who denies Lewis's claim, and mine by implication, that conventions are a species of norm. He offers an argument against any approach where the existence of a social norm is dependent on regular observance of that norm, what I have called 'general conformity'. Southwood gives the example of a hypothetical analysis of the convention among Moldovans not to urinate in swimming pools. It may be that every Moldovan prefers not having anybody urinating in swimming pools, and expects everybody to behave this way similarly—and yet Moldovans frequently urinate in swimming pools. This is meant to undermine the thought that a Lewisian convention is a type of norm, since here would be the structure of expectations underpinned by preferences that Lewis analyses, without the behavioural regularity it is supposed to be an analysis of.[39]

However, Southwood is trading on a peculiar feature of his example. The problem is that for the Moldovans there seems to be nothing at stake in whether they urinate in the pool or not. For there to be something at stake, the Moldovans must at the very least be able to distinguish

[39] Nicholas Southwood, "The Authority of Social Norms," in *New Waves in Metaethics*, ed. Michael Brady (Houndmills: Palgrave Macmillan, 2011), 235-38.

between the state of affairs where nobody urinates in swimming pools and the one where almost everybody does. But *ex hypothesi* they can't. Since they can't usually distinguish between the no-urination and widespread-urination outcomes, their preference for the latter over the former is never brought to bear. Since the preference doesn't come into play when they determine how to act, a convention can't gain any traction.

If anything, the fact that among Moldovans it's widespread that they urinate in pools seems to show that the preferences go the other way to how Southwood constructs the case: they seem to prefer urinating in a pool to not doing so, contrary to what the purported convention requires. Whatever the pools are like, they can tell whether they themselves have urinated, and they seem to prefer the outcomes where they have to the ones where they haven't. At most they're in a tragedy-of-the-commons scenario where every Moldovan would prefer to have a sneaky wee which has a small but cumulative effect on the pools, and this structure of preferences leads to everyone indulging in the same until no pristine pools are to be had. But tragedies of the commons aren't conventions.[40] Even this is too dramatic, I think, since the Moldovans seem unperturbed by the state of their pools. What we really have here instead of a convention is pluralistic ignorance among the Moldovans about the existence of a convention: each Moldovan (at least the ones who indulge in urinating in pools) believes falsely that every other Moldovan universally disapproves of urinating in pools, perhaps seeing themselves as one of the small number of exceptions to the rule. This leads to almost everyone being in error about what Moldovans commonly do.[41] But conventions are things that people commonly do, not things that people merely think are commonly done.

Elsewhere Southwood joins with Lina Eriksson to provide another putative counterexample. They consider the example where a group of friends have an arrangement to

[40] Keeping in mind that I mean here limited or Lewisian conventions.
[41] An excellent discussion of pluralistic ignorance is in Cristina Bicchieri, *The Grammar of Society* (Cambridge: Cambridge University Press, 2005). 186-93 passim.

meet each other for lunch once a week at Imelda's Inn, but each of them will only join in this regular meeting if they are not normatively required to do so. Among the preferences that are being satisfied by the coordination is that none of the lunch-goers have a normative demand upon them to go. Thus, this particular convention requires it not being a norm. Thus, it is meant as a counterexample to conventions being norms.[42]

However, the flaw doesn't lie in the convention, it lies in the way the parties see things. Their desire not to be the object of a normative demand is simply misplaced, and is in fact impossible to satisfy. It is as sensible on their part as a man who resents being called a bachelor but refuses to pursue getting married. Their regular behaviour underpinned by their preference to have lunch together and their expectations of the actions of the other parties just is a norm, or is at least behaviourally indistinct from following a norm. In particular, any lunch-goer's refusal to participate in the behavioural regularity will lead to mutual frustration of their desire to lunch together. Insofar as they are under a normative demand to not willy-nilly frustrate their fellows, they are required to respect their arrangement. It isn't a very serious norm, and it most probably has generous allowances for non-observance. But it is a norm, and the lunch-goers' aversion to calling it one is the same pathology as the resolutely unmarried man's aversion to being called a bachelor. There is a wider point to be made here that we do not in general allow individuals to opt out of norms at their say-so.[43]

The parties to the Imelda's Inn convention are making a mistake, but it's a very interesting one. Their knowledge about when and where to meet each other is first-order knowledge about what the convention consists in, but their mistake about the normative force of the convention is in contrast a mistake on the second order, not about the facts of the convention but about the

[42] Nicholas Southwood and Lina Eriksson, "Norms and conventions," *Philosophical Explorations* 14, no. 2: 200-01.

[43] This makes especially immoderate and unconvincing Southwood and Eriksson's claim that any convention could be rendered normatively inert by way of the participants' desire to not be subject to norms. This claim doesn't seem to have made it into the treatment of Imelda's Inn in the book that they would later co-author. Ibid., 13.

properties of those facts. Something similar also occurred in the Moldovan Swimming-Pool case, where everybody is clear about the first-order action-guidance but mistaken about the role this action-guidance plays. The meetings at Imelda's Inn survive the errors in belief of its participants without any real ill effect.[44]

This thread of argument appears again in the book that Southwood and Eriksson have co-authored with Robert Goodin and Geoffrey Brennan (BEGS for short).[45] In the book they recast their case against the normativity of conventions in terms of necessary and sufficient conditions. Under this construal the Moldovans case becomes a counterexample to conventions being necessary for norms, in that there is a norm without a corresponding behavioural regularity (and thus convention).[46] Similarly, Imelda's Inn becomes a counterexample to the sufficiency of conventions for norms, because the convention to meet at Imelda's Inn conditional on it not being a norm to do so is a convention but not a norm.[47]

My response to the BEGS version of these objections is that their arguments turn out not to target a view like mine. My claim isn't that moral norms just are conventions, but that whatever the grounds of morality are, they include conventions. Thus, my view is non-reductive (morality involves conventions along with a body of principles), whereas the view they target is explicitly reductive (morality is exhausted by conventions). My view is that conventions are necessary for handling SUP cases, which are only a subset of moral cases. Accordingly, the BEGS moves against reducing morality to conventions amounts, on my view, to the claim that morality isn't exhausted by SUP cases, which is neither a surprise nor an objection. Also, my view isn't that conventions are sufficient for moral norms, because I take conventions to be supplements to principles, and limited conventions inherit their normative status from the

[44] This observation about how people can follow a convention without knowing the higher-order features of the convention, or even be thoroughly mistaken about these features, is the subject of Chapters 5 and 6.
[45] Geoffrey Brennan et al., *Explaining Norms* (Oxford: Oxford University Press, 2013).
[46] Ibid., 20-21.
[47] Ibid., 18-19.

principles they supplement.

There are two further issues to cover with the objections from Southwood et al, these being their claim that moral norms are practice-independent, and the question of which views exactly are the targets of their objections. I postpone my discussion of these questions to Chapter 2.

2. The Normativity of Limited Conventions

Here I elaborate on the claim that limited conventions are properly normative responses to strategic underdetermination problem (SUP) cases. As I argued in Chapter 1, limited conventions arise when a group of people develop a settled response to an SUP case, and every such response amounts to a limited convention. Here I develop this point by illustrating how, when conforming to the limited convention, you conform to the principle that limits it. Since in SUP cases there isn't any other way to conform to the principle than through the appropriate convention, this cements the standing of conventions in our moral lives and in our normative practices more generally.

It has been doubted that conventions could play the kind of role I assign to them, and instead could only have an incidental or instrumental relationship to norms, especially moral norms. Brennan, Eriksson, Goodin, and Southwood in their recent joint work have articulated this concern. They give two separate lines of argument against a view like mine. Firstly, they argue that social norms and moral norms never overlap. This is because they argue that social norms are always at least in part *practice-dependent*, in which they mean that the content of the norm varies at least in part with social practices. In contrast, they argue that moral norms are entirely *practice-independent*, meaning that no social practice can influence the content of a moral norm. Since conventions are meant to be social norms, by way of requiring conformity to a continent social practice, this would preclude them having any moral import. The second line of argument they pursue is that conventions can't be norms at all, not even social norms. They claim that conventions are orthogonal to norms, such that the two are fundamentally different and unrelated categories. In Chapter 1 I addressed their case against the normativity of conventions; here I address their arguments against the practice-dependence of at least some

moral norms.

My strategy here is to concentrate on the purposes of actions, and then to show how conventions feature among these purposes, to the effect that if we want to achieve those purposes we need to conform to the relevant convention. This is because some purposes are going to be vulnerable to the *strategic underdetermination problem* (SUP). In order to achieve this purpose we need to be able to predict what the other parties in our situation will do, but we cannot do so just by depending on the shared principles that are meant to constrain their actions. As I argued in Chapter 1, whenever we have a settled response to such SUP cases, it is because we have a *limited convention* for handling that SUP case. A limited convention is a Lewisian convention that coordinates a population in a way that they arrive at a *benign outcome*. Benign outcomes are those where individuals manage to conform to their principles at least as well as they could in any other available outcome in the situation in question. So, insofar as our moral norms have purposes, and these purposes are vulnerable to the SUP, we will need limited conventions to achieve that purpose. This means that those moral norms will be at least partially practice-dependant, meaning that the argument by Brennan, Eriksson, Goodin, and Southwood against the practice-dependence of moral norms fails. That is what I set out to establish in this chapter.

Here I mean 'purpose' in the impersonal sense, referring to what the action is usually able to accomplish and where accomplishing that purpose plays an important role in the aetiology of that action. This is the sense of 'purpose' commonplace in biology and biologically-informed philosophy, such as the work of Ruth Millikan that I make extensive use of below. I exploit the fact that most actions have a variety of effects which may serve a variety of different purposes at the same time. This counts also for the actions that result when the members of a community conform to a convention. We can arrange these purposes on a scale of how distal they are: broadly, the more other steps are required for the purpose to be achieved by the action, the

more distal the purpose is. In some interesting cases we can use this scale to identify a telescoping series of nested purposes, all served by one and the same action.

For instance, consider the act of bringing a gift of wine along if you are invited to dinner. Let us suppose that we're in a community where this is an established convention. Bringing a bottle of wine to dinner is often taken to have the purpose of expressing your appreciation for the favour you were shown by being invited to dinner, and also that you appreciate the principles of hospitality and good company that is meant to be recognised by both you and your host. We can arrange these into a telescoping series: the most distal purpose is 'conforming to the relevant principles hospitality and good company', more proximate than that the purpose of 'showing the right appreciation for the favour you have been shown', more proximate than that 'contributing towards a pleasant dinner', and then most proximately 'conforming to the convention to bring wine when invited to dinner'.

We can describe in general how conventions (and other social regularities) feature in making a norm practice-dependent in SUP cases: by featuring in a telescoping series of nested purposes where one of the more distal purposes is 'conforming to the given moral norm'. Then the most proximate purpose will be 'conforming to some instance of action-guidance', then more distally the purpose of 'conforming to the regularity which results from everybody in the recurring situation repeatedly conforming to that action-guidance', then we have the purpose 'conform to the limited convention that selects that regularity in the given SUP case', then we have the more distal purpose 'bring about the benign outcome', and most distally the purpose 'conforming to the relevant principle'. To conform to a convention is to conform to every stage of this telescoping series.

In §I I provide a descriptive framework for picking out the features of regularities that pertain to action-guidance. In §II I describe the multi-faceted nature of regularities, such that there is a telescoping series of nested purposes that are achieved by conforming to a given

regularity. In §III I use this framework to address the objections of Brennan, Eriksson, Goodin, and Southwood, pressing the point that their failure to appreciate the possibility of practice-dependent moral norms arises from their failure to consider the multi-faceted nature of regularities.

IV. Switching the focus from conventions to regularities

In developing my view, I find it useful and enlightening to give a careful treatment of how regularities relate to conventions, and ultimately to move the focus of our theory from conventions to regularities. This is because I am concerned with action-guidance, and the venue for determinate action-guidance in recurring strategic situations is regularities. This is in turn because the action-guidance that conventions provide comes from the constituent expectations: when the given situation arises, the expectation is that each party will respectively act as expected, and your action-guidance is to yourself act in the way you are expected to. This is not to say that a convention is exhausted by a set of expectations, as I stressed in Chapter 1, but the action-guidance that conventions provide comes from its constituent expectations. The arbitrariness of conventional action-guidance comes from the fact that while there may be many different possible regularities in response to a given case, it is once a particular regularity has been selected that the action-guidance becomes determinate.

By 'regularity' I mean what Lewis meant by 'regularity': when the parties to a recurring situation repeatedly implement a particular set of strategies such that the respective outcome eventuates.[1] For Lewis, a convention is the arbitrary selection of one regularity out of a range of available ones. When people conform to a convention, they conform to the regularity that is conventionally selected.

[1] This involves a set of strategies because the regularity may consist of different actions by different parties. For example, in the recurring situation of people making phone calls that may drop out, there are two strategies: one person reinitiates the call, and the other waits for the reinitiation.

vii. The strategic and individual perspectives

We can distinguish two different perspectives on a convention. Firstly, we can identify a convention with the expectations that lead to people conforming to a particular regularity rather than one of the other possible ones. So, regularities are identified with the conforming behaviour across all the parties, and this behaviour is the object of the expectations that characterise a convention. I call this the *individual perspective*. Secondly, in his analysis, Lewis also identifies a regularity with a certain kind of equilibrium in the social situation, which he analyses in game-theoretic terms. The game-theoretic analysis identifies a response in a strategic situation when the outcomes of their action depend also on what other people do, and the regularity is the repeating course of action that fills that strategic role. I call this the *strategic perspective* on the convention.

The link between the two perspectives is the functional identity between an outcome described as a coordination equilibrium (under the strategic perspective) and that outcome described as the object of the relevant expectations (under the individual perspective).[2] So, to give the dual-mode description: under the strategic perspective, we have the overall strategic situation with the strategies available to individuals leading to a range of outcomes, each outcome offering some given benefit to each of the individuals involved; this situation has identifiable coordination equilibria, which are occupied by regularities. Under the individual perspective, the regularities in turn are constituted by determinate expectations about how the individuals in the recurring situation will act, and these expectations amount to action-guidance. From the strategic perspective, conventions are the arbitrarily selected regularities out of a range that allow for mutually-beneficial coordination.

For many purposes, including crucially Lewis's purpose of providing an account of the conventionality of language, the strategic perspective is especially important. But the strategic

[2] I postpone an extended treatment of functional identification till Chapter 4, where I make extensive use of it.

perspective doesn't exhaust what there is to say about conventions. In particular, determinate action-guidance is a feature of the individual perspective: to say that it is expected of an individual to do such-and-such isn't to report a feature of the strategic situation, but instead to highlight one constituent of the structure of expectations that turns out to play the role of being a regularity in the strategic situation. Normativity pertains to what we should do, and it is the expectations that tell us what we should do. In contrast, the strategic perspective tells us what general properties our actions should have, not what those actions determinately should be. So, to understand the normativity of conventions we should look to the individual perspective. While Lewis doesn't distinguish between the two perspectives, when he gives his own treatment of the normativity of conventions, it revolves about how individuals relate to each other in some determinate situation when held to some determinate expectations, and not to the general properties that the strategic situation imposes on them.[3]

viii. Regularities predominate in the individual perspective

From the individual perspective, in the first instance you deal with regularities, and only derivatively (if at all) with conventions as strategic arrangements. Lewis never makes a distinction like I do between the individual and strategic perspectives, and jumps between these perspectives without comment.[4] But it turns out that we need such a distinction. It is by now standardly accepted in the literature that, *contra* Lewis, individuals can participate in a convention without knowing that they are doing so. The spur for this is an argument by Tyler Burge. Consider the following example: an isolated community is monolingual and has little or no contact with other communities that speak different languages; they take their language to be the sole proper language. Insofar as they are aware of other languages, they consider them to

[3] Lewis, *Convention*: 97-100.

[4] The closest he comes is in his claim that the rules of language are 'tacit conventions', which don't seem to require an understanding of the strategic role the rule plays and only an appreciation for the relevant expectations. He doesn't develop this point. Ibid., 103-04.

be aberrations or just simply mistaken. The fact that the members of this hypothetical community don't know that there are other languages available that would serve as well as their own (they don't know because they hold a false contrary view) doesn't change the fact that their language is conventional like any other. Therefore Lewis's requirement that individuals know that what they are conforming to is a convention for it to count as a convention is too strict.[5]

Burge's point is unsurprising if you distinguish between two different perspectives on conventions, as I have in §I.i. The claim then just becomes that the features that the members of his hypothetical community knows by way of speaking the language fall under a different mode of presentation than features such as 'other candidate regularities are available'. The action-guidance and expectations about how other members of the community will fall under the mode of presentation I've identified as the individual perspective, and the relationship of the regularity they participate in and other regularities falls under the strategic perspective. It's as unsurprising that someone can know the convention from the individual perspective and not from the strategic as it is that they can attest to knowing about Mark Twain but not knowing about Samuel Clemens.

What is pertinent here is that if you engage with the convention by way of the action-guidance they provide—in the language case, following the grammar and using the vocabulary—you are engaging with conventions from the individual perspective, not the strategic. That is, you're engaging with the conventions in the first instance as regularities, and not automatically as the strategic situation wherein these regularities feature. This isn't to say that engaging with conventions from the individual perspective precludes also doing so from the strategic. But these modes of presentation do come apart. And when you engage with conventions in their presentation as collections of action-guidance, it is the individual

[5] Burge, "On knowledge and convention."

perspective and regularities that predominate.

ix. Norms are basic to regularities

In this section I give a brief argument for why norms are a constituent feature of regularities as they feature in conventions. To do so I present a spectrum of types of recurring situation, ranging from those where there is no prospect of norms influencing the regularity on the weakest end to the regularity depending on norms being in effect on the strongest end. I then argue that regularities as they feature in conventions are at the strongest end of that spectrum, and thus have norms as a constituent feature. I'll introduce these three different types by using the example of someone regularly using one of these recurring situations as prompts to set their clock:

I. You set your clock by the position of the sun.

II. You set your clock by Immanuel Kant passing by during his daily walk.

III. You set your clock by the ringing of the town's clock tower at noon.

Type I regularities stand apart by not being social at all, since there is no prospect of interpersonal interaction. There is no question of cooperating with the sun. What happens instead is that you're matching your behaviour to some impersonal phenomenon.

Types II and III are both at least potentially interpersonal and in that respect differ from Type I. Since the party you are matching your behaviour with is another person, they are themselves sensitive to much the same kind of interpersonal interactions as you are. The examples were chosen specifically because the interaction is, as specified, entirely one way—you are responding to Kant's walk and the work of people maintaining the clock tower, while they aren't responding to you. The point is that the phenomenon you are matching your behaviour to is the kind of thing that is sensitive to practical reasoning: different ways of deliberating on the relevant means and ends are liable to result in different actions on their part.

What is interesting here isn't the fact that the other party is open to manipulation. In other

non-personal cases we can sometimes also manipulate the phenomenon we ware matching our behaviour with—for instance, we can orient ourselves with a compass, and compasses can be manipulated by magnets. There are two distinct types of openness to manipulation at work here.

The distinction between Type I and Types II and III cases is the same distinction as Grice makes between natural meaning (as in 'the sun being at its apex means it is noon') and non-natural meaning (as in 'Kant passing on his afternoon walk means it's 3:11pm' or 'it being daytime and the clock striking three times means it's 3pm).[6] It is also the same as the distinction between different kinds of expectations we can have, both described by 'should' and cognate terms in English: descriptive instances, like 'the sun should be directly overhead at noon'; and normative instances like 'the clock-keeper should ring the bells at noon' when describing the obligations of the role. I use this distinction between normative and descriptive expectations to distinguish between Types II and III regularities.

Many examples of normative expectations are also instances of descriptive expectations: when the norm is being followed, both the normative and descriptive expectations regarding the norm will be satisfied.[7] So, in most circumstances we have both the normative and descriptive expectation that the clock-keeper will ring the bells at noon. However, not even Immanuel Kant could be thought to have a duty to be so regular with his daily walk that his neighbours could set their clocks by it. Therefore the distinguishing feature of Type III recurring situations is that there are normative expectations in addition to descriptive ones. So, just as types II and III cases are united in involving personal parties to the recurring situation, types I and II are alike in involving only descriptive expectations and not normative expectations.

For the rest of the chapter and for the thesis from this point onwards, I will only use

[6] Lewis discusses this relation between his view and Grice's at Lewis, *Convention*: 125-29.
[7] If the norm is being flouted, then this will result only in (defeated) normative expectations.

'regularity' to apply to type III recurring situations. Type I cases aren't of interest because non-personal phenomena can't properly be part of a convention, since they can't regulate their behaviour to match with contingent expectations. Type II cases aren't of interest because, while they involve other people, those others feature in the same way that non-personal phenomena do. This has the result that normativity is a built-in feature of regularities in conventions.

x. The determinative and the epistemic orders

To say that normativity is a basic feature of regularities may come off as strange. For one thing, it reverses what is often taken to be the order of priority between recurring situations and normativity. It is often presumed that certain kinds of norms are the result of regularities, whereas I instead take norms to be a constituent feature of those regularities. Below I offer a distinction between different ways we can understand the order of explanation, one which vindicates the usual understanding, and a further one which vindicates my claim that norms are basic to regularities.

My claim is that there is an important sense in which regularities supplement the principles of our moral reasoning, to the effect that conforming to those regularities also is to conform to the respective principles. So, on my view there isn't some sort of practical hierarchy where people first conform to the principles and only later to the regularities, nor where there are regularities that then produce principles. Neither practically precedes the other. What is there is what I call a *determinative order*, where principles limit the appropriate range of regularities. But the point is that we don't learn or implement principles without first learning and implementing the relevant regularities. That is, there is also an *epistemic order*, regarding the sequence by which we learn what we should do.

The usual description of the relation between principles and conventions is in terms of what I have called the determinative order, where principles have priority and then conventions arise later. For limited conventions I endorse this same determinative order, since the

conventions are formed with respect to benign outcomes, and these are dependent on the principles. But the determinative order by no means exhausts what should be said about the relation between conventions and principles. My claim here is that the usual epistemic order is that, in the first instance, you learn some particular piece of action-guidance—do say please and thank you, don't chat back to your elders—where this action-guidance is very often the subject of a convention, and only afterward do you learn the principle behind it—in this case, to show respect to your elders.

I don't claim that in general we can only come to know principles by way of first learning particular pieces of action-guidance. I do point to the fact that very often that is patently how we do learn principles. This also ties into the fact that there are very many cases where we only know particular pieces of action-guidance and not the underlying principles. If it was otherwise there would be little call for moral philosophy, and similarly appeals to accept a general and contentious principle on the strength of an intuitive acceptance of a particular case would be unintelligible. As a further illustration, consider how in the Platonic dialogues Socrates often drives his interlocutors to acknowledge that while they can give examples of some principle, they can't give a description of the principle itself.

In the previous chapter I gave an analysis of how in SUP cases any attempt to do what the principles require is going to refer to a convention if there is a settled response to that case. This has the consequence that for the principle to offer any effective action-guidance in an SUP case, there needs to be a convention that makes the demands of the principle determinate, and thereby practicable. This means that in the epistemic order for knowing what to do in SUP cases, conventions often come before principles. So, my proposal isn't a reversal in the usual order between conventions and principles; instead it distinguishes between two different orders, and shows how conventions precede principles in one important dimension.

V. Regularities can do more than one thing at once

In §I.iv I divided up a regularity into different constituents and highlighted how we can look at these constituents in different ways. This is an instance of a more general point: that we can apply different levels of description to a regularity, which we can have different relations to. In making this point, I am drawing on the work on rules by Ruth Millikan,[8] and work on surface and deep conventions by Andrei Marmor.[9] Also relevant is Elizabeth Anscombe's observation that an action only counts as intentional under a description.[10] There are two points to address here. The first is to describe how regularities do more than one thing at a time, meaning that an attempt to characterise a regularity by some exclusive purpose is likely to misdescribe it. The second point is that the multi-faceted nature of a regularity isn't obvious to those who conform to it just because they conform to it. That is, someone can conform to a regularity under one description—as performing some given purpose—and yet fail to grasp that the regularity performs some other purpose as well.

xi. Millikan on nested purposes for a regularity

Millikan has in multiple venues characterised conventions and rules as a process paired with a telescoping series of nested purposes which are fulfilled by the process in question. The same goes, presumably, for similar normatively loaded phenomena, including social phenomena. To give her example: when a circus poodle learns to ride a bicycle, there is a range of purposes the dog serves this way: the dog does as it has been trained, it pleases its owner, it gets fed, it secures its survival.[11] Since Millikan doesn't give this approach a name, I'll call this the *nested*

[8] Ruth Garrett Millikan, "Truth, rules, hoverflies, and the Kripke-Wittgenstein paradox," *Philosophical Review* 99, no. 3 (1990).
[9] Marmor, *Social Conventions*..
[10] G. E. M. Anscombe, *Intention* (Cambridge, MA: Harvard University Press, 1957).
[11] Millikan, "Truth, rules, hoverflies," 341.. Millikan goes on to claim that all rules and conventions bottom out in biological purposes, "Language conventions made simple," *Journal of Philosophy* 95, no. 4 (1998). I mean to take over Millikan's model of what I call 'nested purposes', but on my presentation it is an open question what the most distal purpose is. I also here don't commit to the thought that there is a privileged purpose, unlike how (in one plausible reading) Millikan wants to prioritise biological purposes. I would find it unsurprising if

purposes model.[12] Placing regularities within this model highlights that there isn't just one action by which we should characterise the regularity. In Millikan's example the one and the same act—the poodle riding a bicycle—performs all of the purposes named above. It is vital to her task that all of the purposes in the telescoping series are performed by the same action, because it is the existence of such a telescoping series which allows the action to become a regularity. To use her example, it wouldn't do the circus dog any particular good to merely by happenstance ride a bicycle sometimes. The point is that the dog must do so in an environment where riding the bicycle is an effective way to also fulfil the other ends in the telescoping series: that is, to please the owner and to get fed.[13] These are distinct tasks that have distinct satisfaction conditions. For instance, they have different relevant counterfactuals. For example, if the circus dog was a usual pet it would please its owner and get fed by offering companionship; if the dog was feral would get fed by scavenging or hunting, without reference to an owner or some task. The circus-dog and companion-dog series diverge at the step of how to please their handler; the feral-dog series diverges from both of these by not having the end of being fed attained by means of pleasing a handler. So, whereas the circus dog and the companion dog both have 'please your handler' as a purpose nested in their more proximate behaviour ('ride the bicycle' and 'offer companionship'), the feral dog doesn't. Of course, any dog is likely to offer companionship or be able to do so in very broad circumstances, sociable animals that they are, but it's the companion dog for whom this has the nested purpose 'please

very many of our established behaviours make reference to our biological natures, but I don't see why this level of description is meant to be privileged, not least because the description at this level is so schematic and lacking in particulars.

[12] Her 'natural conventions' would be an example of nested purposes, but not the only example; the regularity described in her hoverflies example isn't a natural convention, for instance, because while regular across the population it isn't so because of social prompts or pressures.

[13] Millikan discusses at length and at various venues, too many to name here, what the standard of effectiveness in play is here, and settles on it being the case that there are enough successes to allow that process to self-replicate. In her hoverfly example, for instance, she notes that only a small minority of cases of male hoverflies following the 'hoverfly rule' lead to them intercepting females using their spot-tracking method, but enough do to allow a large enough rate of reproduction to secure the survival of that species and its mating behaviour. Millikan, "Truth, rules, hoverflies," 330-35.

your handler', and not the circus or feral dogs.

This point extends further, such that multiple telescoping series are involved. For instance, Millikan's circus dog stands in more than one relation towards its handler, each of which has its own telescoping series of nested purposes, which happens to overlap in the one behaviour of riding a bicycle. In the relationship the dog has to the trainer as a source of food, the dog does what it has been conditioned to do to receive food—in this case, ride a bicycle. The most distal purpose in this series is being fed and staying alive. In the relationship the dog has with the trainer as a member and leader of its group, the dog does what it sees as tending to their relationship and tries to please the trainer—in this case, ride a bicycle. The most distal purpose in this series is having a tight social unit. Despite these telescoping series being different, they can overlap at the same action.

xii. Marmor on surface and deep conventions

Though Millikan uses nested purposes for her account of conventions as well, for my purposes the presentation by Andrei Marmor is more relevant. This is because Marmor's view shares features with both Millikan's and Lewis's, and more directly deals with the kind of large-scale social coordination that is the ultimate concern of this thesis. Marmor defends a view where there are two different levels of conventions at work that are strikingly different to each other without threatening their status as conventional. In broad outlines, he describes these as comprising *surface* and *deep conventions*; when giving a more detailed account of the specific cases of language and law he introduces *coordinative* and *constitutive conventions* as examples of surface and deep conventions respectively.[14]

[14] Marmor's view isn't uncontested, and there are arguments that the law in particular doesn't fit his view. This needn't concern me, because I am using Marmor of an indication of something like nested purposes in the conventionalist literature. Whether the law is actually like this or not, Marmor's view is an intelligible example of the kind of view I want to defend here. C.f. Marco Goldoni, "Multilayered Legal Conventionalism and the Normativity of Law," in *The Normative Dimension of Law*, ed. S Bertea and G Pavlakos (Oxford: Hart Publishing).

Marmor offers as a first approximation of an account of the difference between deep and surface conventions that there are often differences between how explicit a convention is in the mind of the people following it. I'll say why this can't be the final story in §III.iii, but it will do for introducing the distinction.[15] Surface conventions are the kind of thing that someone following them would normally be aware of: speaking a particular language, driving on a particular side of the road, playing a particular version of poker, and so on. Deep conventions are the kind of thing that you could quite possibly follow without being explicitly aware of doing so: speaking a language with a modal construction for the future tense, being part of a regulated system of road rules, playing a betting game with imperfect information, and so on. These deep conventions are conventional in the sense that they could have been different and are kept in place by contingent social practices. However Marmor stresses that they are often not conventional in the sense that they are neither the product of, nor obviously dependent on, anything like an agreement between identifiable individuals. So, on the surface we have coordinative conventions that guide people to some regularity in behaviour, whereas deep below these are constitutive conventions that mark out domains in which people can go on to coordinate.

I want to highlight that Marmor's account of deep and surface conventions exploit telescoping series of nested purposes.[16] The best example of this is the legal theory to which Marmor applies his fully-fledged theory, giving a conventionalist analysis of HLA Hart's primary and secondary rules of law.[17] The primary rules are the particular laws that make up a legal system. The secondary rules are the background principles that Hart identify as giving structure to the legal system—about how particular laws are to be formed and interpreted and

[15] While Marmor appeals to this as a motivation, it can't serve as a way to separate surface from deep conventions, because of Burge-type considerations. The extent to which individuals may fail to grasp that they participate in conventions, even what Marmor would consider surface conventions, is the topic of Chapter 5.
[16] This presages the 'nested conventions' I describe and use in Chapter 4.
[17] H. L. A. Hart, *The Concept of Law* (Oxford: Oxford University Press); Marmor, *Social Conventions*: 155-75.

relate to each other, and so on. These constitute two distinct normative levels, both of which are usually taken to be conventional.[18] Take the 'rule of recognition' for example, which describes the circumstances under which courts should recognise a primary rule (the laws on the statute books). Hart identifies it as one of the secondary rules that mark out the domain and methods of legal practice for the community (by way of the judges appointed by the community who take up that rule). On top of these are the primary rules, the particular laws which are the objects of surface, coordinative conventions, that address the more proximate end of having a law to see to such-and-such circumstance. To conform to one of these laws is also to conform to the relevant rule of recognition—it is to recognise laws that meet the standards settled by the more distal convention. The same goes for the other secondary rules: to conform to any primary rule (a particular law) is also to conform to the rule of change and the rule of adjudication.[19]

Two further features of the account need to be noted. Firstly, Marmor emphasises that it isn't the convention that constitutes the value, but instead the value comes from following the convention.[20] I have made the same point by highlighting how the purpose of a limited convention is to bring out a benign outcome.[21] Secondly, Marmor discusses how there can be many different layers of deep conventions. This is in keeping with what I've discussed about Millikan's view above, and will discuss in relation to Anscombe's below.

[18] Marmor notes this view is widespread, but not uncontested. See *Social Conventions*: 156 n. 6.

[19] I don't want to engage with conventionalism in the philosophy of law, because it is the subject of an enormous literature and this isn't the venue to engage with it. I am concerned with conventions in action-guidance and morality in particular; conventions in the law is a related but distinct field. Applying limited conventions to the legal case would require another work at least as large as the current one, with different interlocutors and views than the ones I engage with here. Similarly, I don't engage with Marmor's claim that surface but not deep conventions are liable to being made into institutional rules (and thus would no longer be straightforwardly conventional). This isn't the venue for me to discuss the institutionalisation of conventions. See ibid., 50-52..

[20] Ibid., 37-38.

[21] The evaluation of the outcome needn't be restricted just to a state of affairs, as consequentialists hold. It can of course include how the participants came to that outcome, e.g. how we distinguish between someone one person giving another something as a gift, and them giving it because they have been blackmailed into doing so. In Chapter 4 I give an extensive discussion of how the psychology of individuals can be a part of evaluating outcomes.

xiii. Regularities aren't transparent

In the course of the chapter I have already given indications that we shouldn't expect people to know everything pertinent about a convention just because they conform to it. Here I press that point a bit further, with reference to Elizabeth Anscombe's work on how actions are intentional only under a description. This is a consequence of my readings of both Millikan and Marmor's views, but Anscombe is more explicit on this point, and her work is more widely known.

Like Millikan and Marmor, Anscombe has a view on action that highlights that actions achieve many ends at the same time. Anscombe presses the fact that because of the way actions are situated in particular circumstances, there isn't any kind of regular relation between doing a particular thing and why you do it. The circumstances mean that to behave in a given way has a multitude of effects, and intentionally performing an action requires having only one of those effects in mind. By parity of reasoning, and as discussed in §II.iii, a given effect also matches with an attitude you have only under a description. This means that the fact that individuals characteristically have different attitudes when considering some object of action doesn't mean that the relationship between them and that object are characteristically different. The characteristic attitudes that individuals can have towards the same object that they are in the same relationship with can vary too widely for that. The attitudes involved towards different purposes in the same telescoping series anchored in the same action can vary independently of each other. For instance, someone can tend their garden both as a way of beautifying their home and as a form of relaxation. If they stop caring about beautifying the home (say, they will soon move to a different house), they can nonetheless continue to pursue gardening as a relaxing pastime.

Similarly, someone can perform an action under one description, and be ignorant about its having an effect not captured under that description, despite that action also having this further effect. For instance, 'kneading bread dough so it will rise' is a description under which bakers characteristically perform it. But kneading the dough allows strands of gluten to develop, which

is a different effect than its rising when baked (as many of us have sadly discovered, there are many reasons bread will fail to rise when baked, not exhausted by the gluten not being developed). The majority of bakers don't know of this effect, and nobody knew about it till relatively recently, despite people regularly succeeding at baking bread for thousands and thousands of years.[22] So, that we have a particular attitude towards one effect of our action—one purpose in the telescoping series—doesn't mean that we have the same or commensurate attitudes towards the other effects of that same action.

VI. Practice-dependent moral norms

There has been a recent comprehensive treatment of norms co-authored by Geoffrey Brennan, Lina Eriksson, Robert Goodin, and Nic Southwood.[23] Because it seems the authors are listed alphabetically rather than according to the extent of their contribution to the work, I refer to the authors by the acronym BEGS.[24] They mean their account to cover norms in their full generality as they appear in the social sciences, with different but related accounts of moral, social, and formal norms. Here I concentrate on their treatment of moral and social norms, especially about how BEGS relate these two domains; formal norms don't enter into my discussion.

As I said in the beginning of the chapter, BEGS have an extended treatment of what moral vs social vs formal norms are meant to consist in, and their interest range far wider than mine. They defend at length a view about how to differentiate these norms from each other, including how these views have different groundings, how different things would count as following and conforming to these norms, and so on. On contrast, I am not here aiming to defend a theory of

[22] Knowing about the effects of your actions in one way but not another is the subject matter of Chapters 5 and 6.

[23] Brennan et al., *Explaining Norms*.

[24] The same practice is followed by Kai Spiekermann in his review of *Explaining Norms*, Kai Spiekermann, "Review of 'Explaining Norms'," *Economics and Philosophy* 31, no. 1. I promise not to make a pun on the line of 'BEGS the question'.

how to define moral and social norms (nor formal norms). Insofar as such a theory matters, I take the requirements of moral principles to be paradigmatic moral norms, and I take norms that regulate strategic behaviour to be paradigmatic social norms. Here I advance nothing more detailed than that these categories overlap, which is commonly taken to be at least a live option,[25] and I target BEGS insofar as they deny this.[26] Accordingly, I have no stake in whether the BEGS view of norms is correct, with two exceptions. Firstly, they deny that conventions are themselves normative, and rather than being seen as paradigmatic social norms should be seen as a distinct, parallel category.[27] I instead hold to the traditional view that conventions are social norms. Secondly, BEGS has moral and social norms be exclusive categories: moral norms are entirely practice-independent, and if something is practice-dependent to some extent, it's a social norm.[28] In contrast, I want to give a wholehearted defence of the claim that there can be practice-dependent moral norms. I address these features of BEGS's account in turn.

xiv. Reducing norms to conventions

When BEGS considers the normative standing of conventions, they do so by consider the view that norms are a kind of social practice, with Lewisian conventions picked out as the most prominent example. This is all well and good, and I have no quarrel with that. But there are two problematic aspects with their analysis. Firstly, they attack Lewisian conventions as an example

[25] For another statement of this as a live option, see Gerald Gaus, "Review of 'Explaining Norms'," *Notre Dame Philosophical Reviews*(2014), http://ndpr.nd.edu/news/explaining-norms/.

[26] My considered view, which is neither presented nor defended in this thesis, is that there is no sharp division between moral and non-moral norms, but that instead the various domains of norms we have—and there are very many such domains, since norms are pervasive throughout human life—bleed into each other. Whereas we can't distinguish sharply between different kinds of norms, we can highlight different kinds of import they have, such that to call something a moral or a social norm is to highlight a particular feature of theirs. That is, I take it that calling something a moral norm or a social norm is a matter of emphasis. What matters for my purposes is that we can identify the social import of conventions, because they regulate interpersonal actions, and we can identify the moral import of acting from principles. So, I take it as a given that if conventions were to play the kind of role in moral decision-making I discuss here, it would count as both a social and a moral norm.

[27] Brennan et al., *Explaining Norms*: 15-21.

[28] Ibid., 72-81.

of a view where norms are reduced to practices, whereas Lewis's view is nonreductive. Secondly, on the basis of this attack they go on to say "that norms and social practices (such as conventions) are crucially different conceptually and functionally, such that it is a serious mistake to assimilate them".[29] Here I argue that instead BEGS's mistaken views on what Lewisian conventions consist in have lead them to the overly quick conclusion that conventions and norms are orthogonal from each other. In this way I wish to preserve the commonsense view that conventions are, in BEGS's terms, to be assimilated under our understanding of (social) norms. That is the view Lewis defended and the one I defend here and in Chapter 1.

It is clear that BEGS believes that conventions can't be normative, but it will do to be clear what they think is at stake. A good amount of ink in the book is spilt discussing why no reduction of norms to conventions is likely to succeed. I won't address this worry, because I think it is misdirected: Lewis doesn't want to reduce norms to conventions, nor do I, nor do most people interested by the role of conventions in norms. By the time they reach their book-length treatment of normativity and conventions, it has become hard to know exactly what views are the target of BEGS's objections. They explicitly exclude views like Margaret Gilbert's which make conventions out to be at least in part normative.[30] This exclusion would presumably also extend to my view, since the product of normative principles along with limited conventions is at least in part normative (though they'll object to my view making some moral norms practice-dependent). They also exclude views that make the grounds of morality out to be a product of convention. In these views, individuals respond to the grounds of moral norms, and these are ultimately derived from contingent social practices (something likely to not be appreciated by many of the individuals involved). They report that they have no quarrel with this latter kind of view, presumably because for their purposes they are interested in how

[29] Ibid., 102.
[30] Ibid., 16.

we get from whatever the grounds of norms are to the norms themselves, rather than in the separate question of where those grounds come from. [31]

This is all well and good, but what views are left for BEGS to attack? They insist that you can't have a reductive account of norms in terms of conventions, and such views are to be found.[32] But they cite these only in passing.[33] The view BEGS cites over and over is Lewis's and views that respond to Lewis's. But Lewis's account isn't reductive. Lewis never makes the claim that norms are a species of convention, or identical with conventions. Instead, he claims that conventions are a species of norms, which of course precludes his account from being reductive.[34] Lewis does believe that conventions are typically sufficient for norms, and BEGS have an argument that attacks that claim directly.[35] But at the end of that they again talk about how this shows a reduction of norms to social practices (conventions included) can't work. This is puzzling, because Lewis's view on conventions precludes such a reduction. This is at best an unhelpful way to discuss what it at issue. The point BEGS hope to make is of more general scope than just Lewisian conventions, but a Lewisian convention is the only practice they explicitly discuss, and there's no discussion of how to extend the point to the more general domain. What is also puzzling is that they admit that their case depends on a contrived situation that is meant to show that while a convention typically leads to a norm, it doesn't do so without exception.[36] Lewis is careful to discuss how the inference from a convention to the existence of a corresponding norm is defeasible—less secure than the already defeasible inference from his list of conditions to the existence of a convention—but still prevalent enough to be of

[31] Ibid., 75 n48.
[32] The survey by Verbeek is useful for identifying these: Verbeek, "Conventions and Moral Norms."
[33] The most extensive treatment of such a view is of Peyton Young's, amounting to a single sentence in the main body of the text. As a footnote to this mention there is citations of two similar views, but they are mixed with ones that don't identify norms with conventions, like Lewis's, Ulmann-Margalit's, and Brennan and Pettit's. Brennan et al., *Explaining Norms*: 15-16.
[34] Lewis, *Convention*: 97.
[35] Brennan et al., *Explaining Norms*: 18-19.
[36] Ibid., 19.

explanatory import.[37] It is unclear just how securely BEGS have targeted Lewis's view, despite it being the opposing view they discuss the most.

Especially troublesome is the fact that in no version of these objections do BEGS or any of their subsets deal with Lewis's positive argument for the normativity of conventions. This matters, because Lewis's argument introduces a feature that isn't handled by their objections. BEGS's objection is that it is possible for there to be a Lewisian convention which has conditional preferences that preclude the existence of a norm.[38] But the normativity of conventions isn't meant to just rest on the fact that the parties to a convention have conditional preferences to conform based ultimately on some strategic arrangement of their desires. What also plays a pivotal role is that the different individuals depend on each other conforming in order to get what they desire, and they have a reasonable expectation both that they can get what they desire and that the other parties will play their part in doing so. This thought has been given a lengthy development by theorists like Seamus Miller and Michael Bratman;[39] Bratman gives this notion the helpful name of 'the interdependence of individual plans'. The idea is that the success of any individual's plan is dependent on the other parties to a strategic situation acting in some specified way. In this pervasive condition we require some way to predict and work with our fellows' plans in order to see our own to fruition. Conventions are meant to be one way to do this. Lewis gives an argument that conventions give rise to justified expectations that our fellows will act in the ways specified by convention, to the effect that if they don't conform and thus scupper our plans, they will have done so in a way worthy of disapprobation.[40] In chapter 1 I elaborated on this point by highlighting how our attempts to follow our principles are similarly vulnerable. Thus, conditional preferences to conform are far

[37] Lewis, *Convention*: 97-98.
[38] I respond to their argument at length in Chapter 1.
[39] Seumas Miller, *Social Action: A Teleological Account* (Cambridge: Cambridge University Press); Michael E. Bratman, *Shared Agency: A Planning Theory of Acting Together* (New York, NY: Oxford University Press).
[40] Lewis, *Convention*: 97-100.

from the only factor in play for the normativity of conventions.

Earlier presentations of this work, by first Southwood and then Southwood and Eriksson, were less irenic than the presentation by BEGS.[41] In the earlier articles the authors are happy to make strong claims about the unsuitability of conventions to the normative domain, with no overt reference to the reduction of norms to conventions.[42] The argument in support of these claims has made it into the book-length treatment, but the claims they are in support of have changed somewhat. Perhaps what has happened is that with the book there has been a shift of focus without a corresponding shift in the arguments, leaving a mismatch between the aims of the work and the arguments given in support.

xv. The argument against practice-dependent moral norms

Separately from their claims that conventions aren't normative, BEGS defend the view that moral norms are practice-independent: what the practices of a community are have no direct bearing on what the moral norms of that community are. The BEGS case against practice-dependent moral norms requires that the grounds of moral judgements exclude practices. They are for this on two fronts. Firstly, they claim that attempts to ground moral norms in social practices are inappropriate on their face. Secondly, they cheerfully admit that there can be practices that occur in a community for the purpose of helping individuals do what their norms require, but they take the moral import of these social practices to be wholly derivative from some practice-independent grounds. I take these to be the operational reasons for why they believe that moral and social norms don't overlap. But before I turn to them, I first wish to discuss BEGS's background reason for this split, which follows from their preferred account of what the grounds of norms are.

[41] Southwood, "The Authority of Social Norms."; "The Moral/Conventional Distinction," *Mind* 120, no. 479; Southwood and Eriksson, "Norms and conventions.".

[42] The responses I gave to BEGS arguments in Chapter 1 work against the Southwood and Southwood and Eriksson versions as well.

BEGS have a substantive view of what constitutes norms (of all three kinds), and this view implies that moral norms are incompatible with being even partly constituted by practices. Their view is that to subscribe to a norm is to display a cluster of normative attitudes. The objects of these attitudes are what determines whether it is a formal norm (if the object is an official sanction), social (if the object includes social practice), or moral (if the object is a moral judgement).[43] Having the attitudes that match with a formal norm have to do with complying with the relevant official sanctions; those regarding social norms involve issues of membership and belonging; those regarding moral norms involve being a decent person. This motivates a sharp distinction between the kinds of norms, because all three are doubly different from each other: a difference in object (the grounds) and a difference in attitude. BEGS don't appeal to the background theory to press their case against practice-dependent norms, but it is part of the picture of norms they present.

The problem for their background reason is something stressed by Anscombe: what purpose someone has in mind when they do something is a matter of description, and there's at least one description available for every nested purpose. The link of attitudes to objects just isn't firm enough or go far enough to do the kind of work BEGS wants it to do. In particular, it isn't exclusive in the way BEGS wants it to be, in that it is commonplace for there to be massive overlaps in the objects of contrary attitudes. All I need for my argument is that different people (or the same person at different times) consider the same action as the object of a social practice and also as the object of a moral norm. BEGS don't have the resources to deny this, given what has been established about how individuals can have different attitudes about the same telescoping series of purposes, as discussed by Anscombe, Millikan, Marmor, and myself.

[43] Relevant here is Southwood making a coarser distinction between moral and conventional interests: Southwood, "The Moral/Conventional Distinction."

Back to the operational reasons. One reason BEGS has for why moral and social norms don't overlap is that social practices and the norms that uphold them just to be the wrong kind of ground for morality.[44] They give a range of examples where they ask the reader to assent that to say, for instance, refraining from murder isn't just a social practice.[45] This is nothing more than a red herring. At most that can show that no moral norm can be entirely grounded in a social practice. And if BEGS restricted themselves to arguing against norms being reducible to practices, that would be fine.[46] But they claim that something is a social norm if it is grounded to any extent in a social practice. The relevant alternative is thus where something is partly a moral norm, and partly a social norm. Certainly, if people are aware of the moral norm they wouldn't stop at the social norm to describe why it is they need to conform to this practice. But this isn't an objection to the practice-dependence of moral norms—it seems to be an illustration of it. This is what you would expect from someone who recognises that this social practice isn't the ground just for a social norm but also in part the ground for a moral norm. BEGS need to provide an independent reason to show why moral and social norms are in competition with each other. And I've argued at length why there is no competition.

More promising is their other suggestion, that social practices may help fill in the content of a norm, but that the social practice merely plays a derivative justificatory role (as they put it), with the underlying moral norm doing the real work. They also stress that it isn't enough for a norm to call upon social practices for it to make it social, otherwise any moral norm which involves, say, linguistic expressions would become a social norm on facile grounds.[47] This is all well and good, but BEGS's conclusion that therefore the norm is practice-independent is too

[44] I don't wish to contest their Grounds View on norms, and am happy to accept it for the sake of argument, but I don't wish to affirm it either. I want to stay irenic, because the purpose of this thesis is to show how any view on the content of morality needs limited conventions if it needs to navigate through SUP cases.
[45] Brennan et al., *Explaining Norms*: 66-75.
[46] It still wouldn't explain why Lewisian conventions are their stalking horse, because (as discussed above) Lewis doesn't reduce norms to conventions, but let us leave that aside for now.
[47] Brennan et al., *Explaining Norms*: 72-75.

quick. The problem is that in SUP cases the practice-independent grounds of a moral norm itself will not suffice for giving determinate action guidance, because these grounds (which are what count as the principles on BEGS's account) are underdetermining. BEGS appeal to, for instance, there being moral norms to obey the law, and then we refer to the practices to know what the laws are. The moral norm that requires you to obey the law is a determinate requirement to conform to some specified social practice. This is determinate in the same way that the duties that people have towards their spouses are determinate, though of course who is whose spouse varies. The same goes for BEGS's other examples: conforming to road rules, not free-riding on social practices, and so on. But settled responses in SUP cases are not like this. In SUP cases there is no determinate answer to what is required of you without a further appeal to (a specific kind of) social practices. Whereas BEGS appeal to cases where the moral judgements point us towards where we should look to see the relevant detail about what we should do, in SUP cases we need the social practices to settle even what it is that we should be looking for. Consider for example the standing of abortion in New Zealand. This is a case where there isn't just an issue of the delivery of a health service, but also a moral judgement about what kind of health services are appropriate to deliver. In New Zealand abortion is strictly speaking illegal, with an exception granted to cases where two doctors (at least one being an obstetrician or gynaecologist) both affirm that continuing the pregnancy will pose serious danger to the woman's physical or mental health. However, in New Zealand physicians have a policy of providing the necessary medical clearance more or less on demand (unless there are medical reasons not to), normally on mental health grounds. Furthermore, in cases where this medical exemption applies the procedure is provided by the clinics under the purview of the District Health Boards (meaning that the government covers the cost), or by licensed private providers, meaning that there is a (not perfectly successful) commitment on the part of the medical system to provide the service. This is more than something that happens a

lot—it is considered to be the best practice amongst New Zealand physicians to offer medical clearance for abortions and to see to their being performed. Not every physician need to do so, but there is a requirement for physicians who object to the practice to tell patients that other doctors can do so, exactly because providing abortion is taken to be the best practice.[48] So, in practice abortion in New Zealand is relatively easily available (leaving aside the real but distinct issue of an individual's access to these health services). What happens here is that we may have expected the status of abortion to be settled by the law, which normally settles these kinds of questions. That would have made access to abortion very restrictive. Instead, the status of abortion is settled by the standards among physicians, exploiting the fact that they can provide a medical exemption, which means that abortion is relatively easily available. This *modus vivendi* has the effect that many New Zealanders are surprised that abortion is strictly speaking illegal, because it is generally accepted and relatively widespread. The result of all this is that we cannot act as if there has been a practice-dependent moral norm that settles what the norms around abortion are—say, that the status of abortion is whatever the law says it is. Instead, the social practices have shaped the standing of abortion, in a way that the road-rules don't settle whether we should try to avoid harming our fellow drivers. So, the moral norms pertaining to abortion in New Zealand are at least partly practice-dependent. This shows that some moral norms are at least partly practice-dependent.

xvi. Conventions are needed for accountability

The last step in making my case is to show that BEGS themselves need practice-dependent moral norms. To do so, I draw upon their positive view about what norms are for: they believe that norms provide the standard against which individuals are held accountable. As they put it:

[48] "Termination of Pregnancy in New Zealand," ed. Best Practice Advocacy Centre New Zealand (Dunedin); Megan Cook, "Abortion," in *Te Ara Encyclopedia of New Zealand*, ed. Manatū Taonga Ministry for Culture and Heritage.

"[w]hat accountability involves is others having a recognized right or entitlement to determine how one is to behave".[49] Here I argue that accountability is itself vulnerable to the SUP, meaning that if there is a settled response to these cases, then it is because of limited conventions. Since accountability is a central feature of BEGS's account, and cuts across all kinds of norms they argue for (moral, social, and formal), if accountability is practice-dependent by way of being at least partly conventional, then BEGS must allow for practice-dependent norms even in the moral case.

That standards of accountability are vulnerable to the SUP follows from two features: that it isn't strongly codified, and that it is at work in strategic situations.[50] To say that accountability isn't strongly codified is to say that there isn't a standard of accountability available which can be applied in every situation and gives wholly determinate results. I don't know of anybody who has tried to provide such a precise and complete description of a standard of accountability—BEGS never do—but at the risk of labouring the point I'll say something about why we shouldn't expect a strong codification.

Accountability in BEGS's use, and in general, is a workaday notion not liable to the kind of precise statement required by strong codification. It is vague what would and would not count as holding someone accountable. Consider, for instance, various much-publicised 'affluenza' cases where people of a high socio-economic class get into legal trouble out of either a disregard for their basic obligations as citizens (such as a teenager in Texas who was guilty of an especially egregious case of drunken driving, killing four people)[51] or an active flouting of those obligations (such as four teenagers in New Zealand who engaged in burglaries for the fun of it, eventually caught after a four-month spree).[52] In these cases these individuals,

[49] Brennan et al., *Explaining Norms*: 36.
[50] More on strong vs weak codification in the Introduction and in Chapter 1.
[51] "Teenager's Sentence in Fatal Drunken-Driving Case Stirs 'Affluenza' Debate", *The New York Times*, 13 December 2013.
[52] "The adrenaline-rush burglars", *The New Zealand Herald*, 13 May 2016.

making use of the extensive resources that come with their high socio-economic standing, were able to extract all the protections the court could offer for defendants in their position, and accordingly very lenient sentences. There has been public outcry about these cases because even though these individuals faced the scrutiny of the courts, undeniably they came off lighter than most other people of lower socio-economic class who are guilty of the same misdeeds. This outcry comes from the difference between what legal trouble most people guilty of those crimes would face, and the amount faced by these affluent individuals—significant but also significantly attenuated. While they have been prosecuted and found guilty by the courts, and thus fulfilled one precisification of what it would take to make them accountable for their crimes, their case fails to fulfil a different and no less salient precisification of what accountability would amount to, treatment and sentencing commensurate with that of other people who commit similar crimes. Therefore, accountability isn't a notion that admits of perfectly clear application, and thus doesn't admit of strong codification.

Clearly BEGS mean accountability to be a strategic notion, otherwise it wouldn't be the kind of thing that needs all of the normative machinery they identify to keep it in place. Any standard of accountability will be interpersonal, since some individual is held accountable to a given standard if the other members of the relevant population treat that individual according to that standard.

If we take as an example the high socio-economic status teenage burglars mentioned above, there's a distribution of labour involved in bringing them to justice. Most clearly it involves the police identifying them as the criminals and arresting them, then the court system prosecuting a case against them and a judge passing sentence; it also involves the members of the community reporting the crimes, suspicious activity, and up to and including turning in the teenagers when they realise that they are the ones responsible. Every individual in this process has a part to play in bringing the teenage burglars to account. In a society like ours, it is neither

appropriate nor in most cases even possible for a single individual to play all the parts from determining that a crime has been committed to bringing some identified individuals to justice. By that token, the various individuals in this process depends on the other parties keeping to the same standard of accountability throughout, because the case passes from one individual to another according to their role in it: it starts with the victims reporting a crime, the police starting an investigation, the parents surrendering the teenagers to the police, the prosecution making a case, and so on. The progress of the case along the stages of this process is paradigmatically an instance of strategic action, because it makes sense to do so conditional on the other parties acting in some specified manner.

So, accountability is strategic, and standards of accountability are vulnerable to underdetermination. This means that accountability is vulnerable to SUP cases. This in turn, means that any settled response to an SUP case involving a standard of accountability is a limited convention, given what was argued in Chapter 1. This means that accountability draws upon limited conventions. To avoid this conclusion, BEGS would need to demonstrate that the argument in Chapter 1 fails, or that contrary to what I've argued here accountability is strongly codifiable, leaving no underdetermination, or that in SUP cases there isn't a settled response about how to apply standards of accountability. Given that I have comprehensively responded to BEGS's concerns against a position like mine, they need to do more to show that my argument fails. No demonstration of the strong codifiability of standards of accountability is forthcoming or likely. It is unappealing to deny that there is a settled response, since standards of accountability won't be able to offer firm guidance (as BEGS supposes) without one. Both the last two options are unmotivated, because by way of limited conventions I have shown how we can respond to SUP cases—like those I've identified for standards of accountability—without requiring strong codification and without us throwing up our hands in despair at instances of underdetermination.

What would a limited convention for accountability look like? Here is a simple, everyday example. Many people enjoy meeting each other at bars, having drinks and conversation. This requires that drinks be bought for all the people participating. A behaviour that is widespread in such circumstances is for the bar-goers to take turns buying each other drinks. Let us imagine a community where this is the established norm. Now consider an individual, Ali, who goes to bars and engages in conversation, but Ali only buys his own drinks. Ali may even go as far as to politely turn down any offers of drinks from others, preferring simply to buy his own. There are many possible reasons Ali may have to not participate in buying rounds. Perhaps he finds that joining others for rounds means he drinks at a rate faster than he would like, perhaps he finds buying rounds overly familiar, perhaps he does it to hide that he is only having non-alcoholic drinks. Whatever the reason may be, the other bar-goers can respond to Ali in different ways. They can hold Ali accountable for flouting the round-buying norm, and hold it against him. Or they chalk it up as a harmless idiosyncrasy of his, in which case there's nothing to hold Ali accountable for. Whatever they settle on, it needs to be a nearly exceptionless response. If they settle on it being fine for Ali to buy his own drinks, then for a different individual Bob to voice disapproval would show Bob to be the one out of step with social norms, not Ali, because Bob is sowing dissention. Similarly, if they settle on Ali being in the wrong on this point, if Carlos doesn't show the appropriate disapproval of Ali's behaviour, then the other bar-goers have reason to disapprove of Carlos in turn as encouraging Ali in his bad behaviour, or just as someone who isn't holding their fellows appropriately to account. This is a strategic case among the bar-goers. It is obvious that the principles of hospitality and good company don't settle how it is the bar-goers should view Ali's aberrant behaviour. So it is an underdetermined one as well. This means that whatever response the bar-goers settle on eventually, whether it be to hold Ali to account for flaunting the norm or not, it would be a convention to do so. And since this is a convention about how to apply a principle, in particular

about how to hold someone accountable to the principles of hospitality and good company, that would be a limited convention regarding standards of accountability.

VII. Conclusion

Here I have elaborated how norms, including moral norms, can be settled by convention and be practice-dependent, and are so in SUP cases. This is because conforming to the limited convention that has been established in an SUP case is just one level of a telescoping series of purposes served by doing so. That telescoping series also includes providing action-guidance and conforming to the principle that is supplemented by the limited convention. I then used this analysis of conventional regularities as multi-faceted and intrinsically normative to address the objections against the normativity of conventions and the practice-dependence of moral norms by Brennan, Eriksson, Goodin, and Southwood.

3. Conventional Authority

It is a commonplace that many authorities are established by convention, but there is no standard way to flesh out this everyday understanding. Here I offer an analysis of *conventional authority* such that a command is genuinely authoritative when conforming to it leads to a particular kind of convention. This is different from the more common suggestion that it is a convention that somebody has authority.

On my account conventional authority is a mechanism that addresses what I call *strategic underdetermination problem* (SUP) cases. This is where individuals share a body of principles, but don't know what would be the best way to follow those principles in a particular situation because the principles underdetermine what they should do. In SUP cases individuals need to be able to predict what their fellows will do in order to be able to determine what they themselves should do. In response to SUP cases we need what I call *limited conventions* which are Lewisian conventions but where we evaluate the different options not based on how they match individual preferences but rather to the extent to which they match the pre-existing principles. This means that the options that individuals may coordinate on are limited to those that are at least minimally compliant with their shared principles. On this account, a command with conventional authority gives an exclusionary reason and should be followed because it establishes the shared expectation that everybody will respond to the SUP case in the specified way, and that response that conforms to their body of principles. If those conditions are met, not following the command undermines everybody's ability to do as their principles require, meaning everybody has an exclusionary and pre-emptive reason to conform to the command.

In support of my analysis I give a conventionalist analysis of parental authority, which is usually not taken to be conventional. Linking the authority of commands to bodies of shared principles allows me also to offer an analysis of how different authorities may overlap, and of

how conventional authority can handle moral variation within a population.

The structure of this chapter is as follows. After a preamble in §I, in §II I argue that a command leading to the forming of a limited convention is sufficient for something being genuinely authoritative. In §III I place my account of conventional authority in context within the literature. In §IV I offer my conventionalist analysis of parental authority, and in §V I discuss overlapping authorities and moral variation within a society.

VIII. Preamble

To forestall confusion, I want to make a clarificatory point right away. Here I offer sufficient conditions for some command to be an instance of justified authority, but not necessary conditions. My account is compatible with there being multiple sources for authority, and conventional authority being only one source amongst many. It would be excessively tedious to continuously stress that I'm offering only sufficient and not necessary conditions for genuine authority, so I ask the reader to keep this feature in mind.

I also want to highlight one unusual feature of my account: I focus on authority command-by-command, rather than individual-by-individual. Most theories of authority try to explain why some particular individual or role can be vested with authority. My account however focusses on what specific commands may carry authority. I propose that for conventional authority at least the commands are primary, and some individual becomes vested with authority derivative on the commands that they typically issue. How we can go from specific commands being genuinely authoritative to individuals becoming vested with authority is the topic of §II.i.[1]

I must also address here a possible mismatch between what commands require and

[1] I will depend on context to keep these two senses separate in what follows. Also, I will use 'authority' in the sense where it is an open question whether that authority is legitimate or merely putative, and explicitly add 'genuine' or some synonym when I want to say not only that a command has been issued, but that those who are subject to it have a pre-emptive reason to follow it.

conventions deliver. Conventions are, at least on Lewis's analysis, about regularities in action across recurring situations. They concern multiple instances of an action repeated over a period of time. But commands very often are one-off events: someone delivers the command, somebody else responds, and that's the end of the matter. Conventional authority then seems to miss a very large part of what we expect from commands and conformity: that we should allow for commands to be authoritative even in one-off cases.

The most important point in response is that the ability of conventions to offer guidance is unaffected by whether they get repeated or not. Since guidance is what is at issue in my analysis, I can cheerfully allow that the convention needn't stretch over multiple occasions. On this point my purposes diverge from Lewis'. Lewis is after a way of establishing co-operation without requiring explicit agreement or pronouncements. For him, the fact that he will only call repeated cases conventional is a stipulation to make the central features of his analysis clear.[2] But of course in the case of someone being subject to commands there is something that can guide the parties: the commands in question. My claim is that these commands play the same epistemic role as a pre-existing structure of expectations do in Lewis's analysis, by giving individuals a way to tell how to reliably navigate through SUP cases. So, the kind of worries Lewis had about how conventions could be established and maintained without an agreement or pronouncement simply doesn't apply in this case.

In any case, while the extension of a once-off command over a range of repeated situations isn't explicit, it is possible. It is commonplace that one-off commands end up shaping behaviour in future cases as well, because these one-off commands create a precedent to be followed in future cases.[3] The existence of 'standing orders' in military contexts, meant to

[2] Something that critics sometimes take him to task for, e.g. Margaret Gilbert, "Agreements, conventions, and language," *Synthese* 54, no. 3..

[3] An extended philosophic treatment of this point can be seen in David Braybrooke's argument for how act-utilitarianism collapses into rule-utilitarianism in non-ideal conditions. Braybrooke, *Utilitarianism*: 14-19.

regulate behaviour over repeated instances by way of the same mechanism as one-off orders, is an indication of how this would go. We can turn any one-off command into a standing order by way of adding 'this command also applies to future instances of this situation'. So, for my purposes there is no deep difference between a one-off command and one that applies to a recurring situation.

IX. A criterion for authoritative commands

I propose what I call the *commands-as-conventions criterion*:

> A command carries genuine authority if a limited convention would result if the people subject to it conform to the command.

I will refer to the above as simply 'the criterion'. It amounts to the claim that commands with genuine authority are a supplement to the moral principles that generally guide the community in question. That is, there is some set of principles that the community subscribes to, which has various consequences about how people in that community should behave; the commands of a legitimate authority is a way to give determinate moral guidance which is not entailed directly by the shared principles, but which is consistent with it, and where it matters that the members of the community are in agreement about how to act.[4]

I will briefly reiterate the relevant background for limited conventions that I presented in Chapter 1. Limited conventions are instances of Lewisian conventions that arise in response to what I call *strategic underdetermination problem* (SUP) cases. The underdetermination in question is where the principles shared by a community are such that there are cases where they only go part of the way towards determining how to respond to the situation. In particular, there are multiple options that remain after the principles dismiss various options as unsuitable, but the principles give no way to choose between them. I call the remaining options *benign*

[4] Since conventions are about repeated behaviour, this criterion may appear not to allow for one-off commands. I will handle this issue later in this section.

outcomes, because they are ones that are not malignant (that is, not determinately worse than another available outcome). The strategic part of the problem is that if you're in an SUP case, to reliably reach a benign outcome you need to be able to predict what your fellows will do. But since the case is underdetermined, you can't use the principles to tell what even a perfectly conscientious person will do. So, the uncertainty individuals face in their own decisions then bleeds over to the decisions of other people, making them uncertain as well.[5]

The way a command becomes authoritative in an SUP case is when it picks out one benign outcome from the underdetermined range and makes it common knowledge that the chosen option is the one to be taken. This removes the uncertainty of SUP cases, which in turn gives the parties reason to follow it. In particular, the subjects have good reasons to follow the commands and the resulting conventions, because they are party to the mutual benefit that comes from conforming, and they would risk everybody's ability to reach that benefit by failing to conform. They also have no good reason not to conform, since the other outcomes they could reach are not determinately better than the one selected by the convention.[6]

The above is not to deny that the different available options will have different distributions of burdens and benefits. My response is that the community's ability to reach any benign outcome at all, and so discharge the demands of their principles, is dependent on their ability to coordinate their efforts. If some attempt to reach one option with its distribution of burdens and benefits while others try to reach a different one, they are going to work at cross purposes. The members of the community need a way to have the necessary confidence in what the other parties will do, so that they can embark on one option rather than another. The commands of some recognized authority will suffice in forming the shared expectation that those subject to it will all act in the way that counts as conforming to the command, and that

[5] I gave a fuller description of SUP cases in Chapter 1.
[6] This is the same argument for the normativity of limited conventions given in Chapter 1, §III.

way of acting also suffices to coordinate towards one of the benign outcomes. This is what I mean when I say that authoritative commands can create conventions that serve as extensions of the shared principles that guide a community.[7]

xvii. Benign arbiters

The criterion that I have provided evaluates instances of authority command-by-command. But typically we take authority to be a property of individuals, and often a property the individual has in virtue of some station that they occupy. For example, the paradigmatic instances of authority are those of a judge presiding over a court case, or a teacher leading a classroom. We don't as a rule try to offer a justification individually for each of the commands issued by these authorities. I don't deny that this is the usual form in which we find authority. To match up my criterion with this observation, I introduce the standard of a *benign arbiter*.

To count as a benign arbiter, an individual has to fulfil two requirements. Firstly, the *salience requirement*: there must be a domain where it is common knowledge that the individual's commands are meant to be respected. This means that there is both a range of issues and a group of people who are subject to the authority on those issues. Whether this is by some personal quality or because they occupy some station to which authority attaches is immaterial. Secondly, the *judgement requirement*: the individual's commands as a matter of fact always select a benign outcome. This means that in cases where the principles of the community in question determine a sole outcome to work towards, the command guides the community to do so, and, in an SUP case, the command picks out one of the range of benign outcomes.

Please note that the notion of a benign arbiter makes no explicit mention of coordination. They are the kinds of things that a community can coordinate towards, and should coordinate

[7] A fuller description of this state of affairs is to be found in Chapter 1, where I also argue that every such response to an SUP case is a Lewisian convention.

towards because of the benefits to one and all of doing so, and the fact that not conforming strips your fellows of these benefits without any countervailing reason. But the command comes first and only afterwards the coordination.

It is also important to note that a benign arbiter isn't some kind of ideal observer, perfectly impartial judge, benevolent archangel, or a superiorly endowed individual of any sort. Those familiar theoretical devices are of an individual in some idealised decision-making position which is meant to indicate that the decisions they make (or would make) are the best choice,or the one we would be best served to imitate, or something of that sort. In contrast, I provide here no specification at all about the type of decision-making position, idealised or otherwise, a benign arbiter would be in. All that matters for my account is the brute success of the benign arbiter in picking out the benign outcomes, however it may arise.

Another reason that there are no specifics about the decision-making position a benign arbiter would be in is because we expect that various individuals meet the criteria in different ways in different domains. The decision-making capacities that make a judge appropriate to preside over a court case—knowledge of the law and precedent, good judgement on the relevance of some piece of evidence or its likely effects on the reasoning of the jury, etc.—are very different from what equips a teacher for their task—sound pedagogic practice, the ability to improvise in response to the students while preserving the integrity of the curriculum, and so on. Thus, on the present view, there is no one set of properties that make an individual likely to make good commands. Accordingly, I concentrate only on whether the commands are actually successful in guiding its subjects to a benign outcome.

xviii. Conventional authority and nested purposes

On my analysis, a command has genuine authority when it helps its subjects to conform to their

principles.[8] My favoured way of articulating the relationship between the command and those principles is in terms of nested purposes, as I discussed in Chapter 2. Conventional authority is an example that illustrates what I have called the multi-faceted nature of regularities, since conforming to a particular command accomplishes all of the nested purposes in a telescoping series. 'Following the command' is just one purpose in this telescoping series. As I describe it, conforming to an instance of conventional authority has the most proximate purpose of doing what is commanded; a more distal purpose of participating in the regularity that arises among those subject to the authority when they comply with it; more distal still is the purpose of securing the outcome that arises from that regularity; and the most distal purpose is conforming to the underlying principles.

X. How conventions provide pre-emptive reasons to conform

To indicate how conventional authority gives pre-emptive reasons to conform, I argue that conventional authority falls under the service conception of authority advanced by Joseph Raz, which is the most prominent account of how authority becomes genuine because of the benefits it gives to those who conform to it.[9]

Raz's service conception of authority is based on three theses, all of which are also satisfied by conventional authority. By way of fulfilling these three theses, conventional authority is shown to be consistent with the service conception of authority. This means that it is a plausible version of theoretical authority, insofar as Raz's service conception is.

First, the *dependence thesis*: that the reasons the authority works from are also the reasons of the subjects. This is satisfied by way of the criterion's demand that commands be consistent

[8] Keeping in mind the proviso that this offers sufficient but not necessary conditions. Also, this claim may seem to be trivialised in the case where the principles include a general principle of obedience to authority. I discuss this later on in my treatment of parental authority, in §IV.i.

[9] Raz, *The Authority of Law*; Michael Sevel, "The Constitution of Authority: A review of Joseph Raz, Between Authority and Interpretation: On the Theory of Law and Practical Reason," *Jurisprudence* 5, no. 2.

with the shared principles, which are the subjects' relevant reasons in the case at hand.

Second, the *pre-emption thesis*: that once the command has been issued, the reasons the subjects have to follow the command trump the reasons they have to act on their own discretion (at least in the preponderance of cases). Conventional authority gives pre-emptive reasons to conform to them because failing to conform would frustrate the resulting limited convention, and thus frustrate your fellows' ability to do what their principles tell them to, since in SUP cases you need limited conventions in order to do what your principles require. I say more about this below.

Third and finally, there is the *normal justification thesis*: that there must be some answer to the question of why there could be binding commands in the first place. On my view, commands with conventional authority are ways to form limited conventions, and conforming to a limited convention is in turn a way of conforming to the underlying principles. Since conformity to the principles is what is required for justification, conventional authority is thus normally justified.

Let us linger on the pre-emption thesis. The reader may worry that at best conventional authority provides reasons to do as commanded, maybe even compelling reasons, but not a kind of reason that trumps other concerns you may have. But this response fails to take seriously that in SUP cases conforming to a limited convention is the only reliable means for conforming to your principles. It isn't that the reason to conform to the command is defeasible because the individual may have some reasons to do something else; if the principles make a requirement on the members of the community in this case, that just means that they have an antecedent obligation to do as the principles require. In SUP cases, if there isn't a limited convention in place then there is no reliable and effective means to discharge this obligation, but it doesn't mean (on most accounts of obligations) that the obligation simply disappears—this is the point behind the notion of a moral remainder. But the commands that carry conventional authority

put a limited convention in place. Conforming to the command and the convention that results is the only way to do what your principles require, and is thus pre-emptive.

The reader may now worry that this standard is too strict, because after all not every principle is itself overriding. It is an important point that principles may themselves allow for quite broad non-compliance. For instance, consider the norms around being a good Samaritan: there is general assent that we have requirements to aid our fellows if they are in distress, but similarly it is commonly held that trying to enforce such a norm is ineffectual. This means that we allow for many instances where someone is faced by a fellow in distress, and yet they aren't expected to help. My response is that in such a case, there can't be the general expectation that the members of the community will conform to the commands, since it is known that the reasons offered by the principle in this case simply isn't that binding. This in turn means that the salience requirement can't be met—there simply isn't anything that counts as the right standing to issue commands in this case, at least not by the standards of conventional authority. And since the salience requirement isn't met, in the case that a principle is permissive enough there isn't anything that will count as a command that carries conventional authority. So, we have good reason to expect to see conventional authority arise only in matters where the principles require quite strict compliance. How strict the compliance would need to be depends on the details of the principles in question, which I leave unsettled in my account.

XI. Parental authority as conventional

We usually consider the authority of a parent over a child to be a paradigmatic instance of authority that is established not by convention but instead by virtue of the *natural relationship* (to coin a term) in play.[10] However, it is also an instance where the parent is in a better place to judge what is good for the child than the child can on their own. This opens to door for an

[10] I refer to the parent and the child in the singular, mainly to be concise. It doesn't influence my analysis at all who exactly the parents and the children are, only what the content and justification of the commands are.

attempt to describe the commands of a parent to a child in terms of theoretical rather than practical authority. Here I endeavour to show how we can give an adequate account of the authority of parental commands using conventional authority, even though parental commands are normally described without any reference to conventions.

There are countless potential harms that threaten a child that a child may not foresee or fail to appreciate, from the discomfort of a sore stomach to the mortal danger of electrocution. The parent is normally a better judge of what would be good for the child to do, and given the natural relationship we expect that the child will obey the parent, at least in general. These indicate that the parent is likely to fulfil respectively the judgement and salience requirements of being a benign arbiter. This goes a long way to establishing how the parent's commands may form limited conventions. To go the rest of the way, I'll discuss first commands that regard the running of the parent's and child's household, and then commands about the child's behaviour more generally.

We must be careful in delineating the extent of the parent's authority, otherwise we will invite confusion about the scope of the conventions their commands bring about. Firstly, the parents are also the people who run the household, and they have a certain amount of authority simply on that account, whether it is over their children or boarders or visiting relatives. An example would be the times at which various things happen inside the household. Secondly, they have the standing to make obligatory or forbidden within the household certain actions that are up to an individual's discretion in the community as a whole. An example of this is whether smoking is permitted at home. Thirdly, and more importantly, the parents have a say in the behaviour of the child and the organisation of the household, but this holds only within the constraints put in place by the community's shared principles. We can see the commands of the parents as being how the child becomes raised into life under the shared principles alongside the other members of the community. This involves learning the conventions that express the

principles within that community (as discussed above), but crucially also includes the fact that cases specific to the household must also be addressed in a way consistent with the principles, even if the principles underdetermine them. There will be a large variety of such cases, most of them innocuous and unremarkable instances of where a household needs to arrange its affairs. We should understand parental authority as the standing the parent has to settle the responses to these cases within the constraints of the community's shared principles. My conventionalist analysis of parental authority is the claim that it is sufficient for a parent to settle these cases by issuing commands that establish conventions regulating the household's behaviour.

The setting of a curfew or bed-time is a domestic example of how a parent's command leads to a limited convention. Each of the different times that a curfew or bed-time would be established at would amount to a different a strategic arrangement, because it involves the coordination of the parent and child around the chosen candidate. It is also a choice which plausibly the shared principles in play won't uniquely determine. It is thus an SUP case—a mild one, buts it will do as an illustration. The response to this particular SUP case serves a variety of purposes, and I'll list three. One of them is that a child needs to have enough rest in order to see to the day's tasks. Then there is the prosaic issue of managing the schedule of the household, which likely falls simply under the parent's authority as the one who runs the household. A further important matter is both the parent and the child having firm expectations about what counts as normal and safe states of affairs. A curfew or bed-time sees to all of these purposes. By having a set time at which the child must be at home or in bed, the parents can establish a regularity in behaviour whereby the child will have the necessary amount of rest. It also allows them to plan their daily schedule against a fixed the strategic backdrop rather than being uncertain about how the other members of the household will act, e.g. they can plan when to get up from bed and what their morning routine would be. Thirdly, having the child be absent past the established curfew can be used as a signal that something is amiss—the child is

either not doing their part to keep to the household's schedule, or (one hopes not) is in actual danger—and marks a determinate cut-off-point where an intervention is called for.

The setting of a curfew or bed-time is thus a clear example of putting in place a structure of expectations to regulate the behaviour of the parties in a way that allows them to engage in effective and reliable strategic decision-making. The natural relationship between parent and child makes the child meet the salience requirement, and their greater maturity and expertise should suffice to meet the success requirement, meaning that the parent can act as a benign arbiter: they can issue a command which, if everybody conforms to it, will lead to a limited convention about the daily schedule of the household. And as for curfews and bed-times, so for the other questions about how the household should be run.

The above doesn't exhaust what parents command their children to do. Some of the commands don't involve the running of the household, but of the child's conduct in general. This becomes a more important role of the child's upbringing the older and more independent the child becomes. When a parent forbids a teenage child from engaging in sexual activity, this is often not a command to not inconvenience the household by way of doing so, or not to do so inside the home, but not to do it at all. This goes outside of the boundaries of household regularities. So how should we understand parental commands that don't work towards regularities in the household? There is the safety consideration, in that the parent can direct the child away from the dangers of sexual activity even when the child does not appreciate them (as is a common issue when a course of action has many attractions but experience with the possible harms is lacking). The most obvious example here is the harm it would do to the child's future prospects if they were to have a child of their own before they are fully established as self-sufficient adults.

There are of course a large variety of strategies for avoidance of these harms that can be put in place, and very many different ones are put in place in different societies, varying from

blanket bans on pre-marital sex (where marriage is seen as the point where an individual is a self-sufficient adult), through to practices that accommodate teenage sexual dalliances up to and including arrangements for dealing with teenage pregnancy in way that isn't injurious to the parties involved. For a matter of this complexity, it is again plausible to consider that which strategy is best isn't uniquely determined by the underlying principles, so we again have an SUP problem, and quite a serious one.

Each of these different arrangements will involve at least the parent and child. But of course, the parent's commands on this score also creates regularities in the behaviour of the child which fit into the practices of the wider community. The child is likely to be uncertain about how exactly the members of the community depend on each other to act in particular ways through their sexual mores and the attendant expectations which allows them to navigate this especially tricky strategic domain. The child is also especially vulnerable to underestimating the import of these regularities in behaviour (through a lack of experience of the harms the conventions secure against, or an improperly high regard for the options the conventions close off, for instance), so the child is given less discretion in their action and expected to conform to the commands of their parent, the adult member of the community best placed to socialise the child into the existing practices. What is at issue is that the community's conventions are the settled responses in the face of the SUP cases that infects any community's attempts to live according to their sincerely held principles. The parent's commands still form a limited convention, in this case not just within the household, but about how the child in particular is to slot into the wider limited conventions found within the wider community. And so too for other non-household-regarding parental commands.

xix. Is this account of parental authority plausible?

Parental authority may be an attractive example for my view, but it is also a much-studied field in its own right, and I need to at least show that the conventionalist analysis is plausible given

the established approaches. While there is a vast literature to do with the relationship of authority between parents and their children, there is comparatively little that directly addresses the moral standing of parental authority. There is a very large literature in developmental psychology on parental styles, regarding how authoritative behaviours on the part of the parents affect the children. However, it says little about the moral standing of the relationship. Instead, it gives empirical guidance on how various styles of parenting actually promote the health (physical and psychological) and good future outcomes of the child. The preponderance of evidence in developmental psychology shows that children benefit from firm but responsive guidance from the parents (the so-called 'authoritative parenting' style), and that attempts to not subject the child to much or any authority do less well (the so-called 'liberal style'), as does an insistence by parents on rigid compliance to their commands (the so-called 'authoritarian parent' style).[11]

Since conventional authority results in definite commands with the expectation that these commands will be followed, it falls within the authoritative or authoritarian styles, not the liberal style. It doesn't require an authoritarian style, and allows for the responsiveness called for by the authoritative style (I cannot make a stronger claim without specifying some determinate content for the principles in question). This is because the content of the command is meant to be responsive to the principles, and of course the principles can allow for responsiveness by parents of the needs of children. You'd hope that principles include at least that much, and can articulate both what features of children parents should be responsive to, for instance by highlighting features of individuals (children included) that are of moral import. It's not just the brute fact that the parent commands something which makes it authoritative, but that the command succeeds in bringing about some genuine moral good, and not in a way that

[11] See the survey in Diana Baumrind, "Authoritative parenting revisited: History and current status," in *Authoritative parenting*, ed. Robert E. Larzelere, Amanda Sheffield Morris, and Amanda W. Harrist (Washington, DC: American Psychological Association).

is determinately worse than another one available. Even if the moral good in question is 'obedience' (as some set of principles may very well allow) a command meant to instil obedience can only have conventional authority if it doesn't neglect the other goods in an avoidable way.

In moral philosophy there is some work done on the standing of parental authority.[12] I want to highlight Robert Noggle's argument that a parent's commands are authoritative on account of the child's abridged participation in a fully developed moral framework. Through leveraging a Rawlsian model of decision-making, Noggle highlights how a child doesn't yet have a particular conception of the good, not having yet built up the store of experience needed to distinguish and identify one substantive conception from another.[13] Because the parent does have the experience and such a conception, they have a fuller extent of agency than the child, and the child is benefitted by having the adult act on their behalf and introducing them into a life conforming to such a conception.[14]

My account is consistent with Noggle's and is more general. Furthermore, my account has an advantage over his. Noggle doesn't address the question about how different conceptions of the good would lead to different distributions of burdens and benefits, with knock-on effects on the decisions individuals can make about what to do. To supplement Rawls's primary goods, which are needed to pursue any conception of the good, he introduces the notion of secondary goods: things which are bad for nobody but only positively good for people who pursue some subset of the possible conceptions of the good. As the parents commit to a particular substantial

[12] An addition to the work discussed here, there is an enormous literature in especially bioethics and to a lesser extent the philosophy of law about who should get to make various decisions regarding things affecting children. This isn't relevant to my analysis, since I am neutral on who it is that fulfils the role of parent as identified here, and there is no supposition that it has to be the biological parent. If circumstances require that somebody else plays the role of guardian, that would make no difference to me.

[13] Robert Noggle, "Special agents: Children's autonomy and parental authority," in *The Moral and Political Status of Children*, ed. David Archard and Colin M. Macleod (Oxford: Oxford University Press).

[14] This last move has been criticised by questioning whether a child would be better off independently deciding on a conception of the good rather than passively being raised into one. Jeffrey Morgan, "Children's rights and the parental authority to instill a specific value system," *Essays in Philosophy* 7, no. 1.

conception of the good, by that token they are also likely to pursue the particular bundle of secondary goods that is in aid of that substantial conception. The problem is that different bundles of secondary goods will benefit different individuals depending on what substantial conception of the good they pursue, and the choice between them is not straightforward. From what Noggle has told us, someone may seem to be harmed (if only by an opportunity cost) if their parent provided them with secondary goods suited to the parent's conception of the good, but not one they end up pursuing once they are capable of choosing a conception of their own.[15] Thus, it is unclear what secondary goods it is best to provide to the child.

My conventionalist analysis avoids the problem by addressing underdetermination head-on. Because my account is built upon the requirement for people to coordinate the ways they follow their sincerely-held principles, we can identify a sufficient reason to accept a particular distribution of benefits and burdens when other ones were available: none of the options could be reliably attained without such coordination, and none of the options are determinately worse than the others. To select a substantial conception of the good is also to highlight a particular bundle of secondary goods, with its accompanying distribution of benefits of burdens. Since conceptions of the good (and bundles of secondary goods) can only be garnered by way of strategic cooperation, whichever substantial conception of the good parents pursue, they need to pursue one of them. By pursuing substantial conception of the good, the parents gain licence to pursue a respective bundle of secondary goods. Since it is fine that parents pursue the substantial conception of the good that they do, by way of it being unavoidable that they pursue some substantial conception of the good, it is also fine for them to provide the bundle of secondary goods that they do.

The above establishes that my account of parental authority is consistent with the relevant empirical literature, and compares favourably to prominent views regarding the moral status of

[15] This is also a major point in Morgan's objection to Noggle's view.

parental authority.

XII. Dealing with moral variation within societies

Discussing the extent of parental authority runs us straight against the very difficult question of how authorities within a community can lead to people being subject to distinct but overlapping moral demands, such as children being subject both to their parents and to the overarching authorities of their community. This is ubiquitous within authority structures, since it is only by exception that an individual has only one master (the proverb to the contrary notwithstanding), but instead is almost always faced with a range of different authorities. Sometimes these authorities are arranged in something like a chain of command, or sometimes with distinct but overlapping domains in which their commands are authoritative. I concentrate on this overlapping-domain case because it seems to me the most difficult to analyse. For a chain of command, there is going to be some order of priority which will decide which command should be followed in case they conflict. As imperfect as this solution may be—and all of us have concrete experience of how chains of command can lead to confusion—for the case of overlapping domains we don't have even that much, since there aren't such priorities to appeal to. A theory that can account for overlapping domains of authority to that case will accordingly be more interesting.

Parental authority is an interesting example of overlapping domains of commands and conventions. It is one thing to have parents raise the children into a single society-wide way of life, in which case there isn't going to be any serious disagreements between households about how the child will be raised to act. It is another if there are substantially divergent views about how households should live and raise their children. This latter case can arise even in the presence of a set of shared principles that all households sincerely subscribe to. It could be that the principles underdetermine the kind of household that subscribes to them—for example, under a principle of concern for the well-being of non-human-animals each household would

reject animal cruelty, but only some of them may end up going to the extra step of adopting vegetarianism or veganism on moral grounds. Or it could be that the principles explicitly allow, even invite, different ways of life amongst its adherents, as many liberal or cosmopolitan societies attempt to do.[16] And if it is allowable for households to have different ways of life, it must also be allowable for them to raise their children into these different ways of life, if we are to take the notion of a way of life at all seriously. Accordingly, we should expect that it is by the parent's authority that the children get raised in a particular way of life, rather than another.[17]

My suggestion is that we can develop a model of nested conventions, where there is a set of wider conventions (covering a larger swathe of cases) that partially determine the content of narrower conventions (dealing with more specific cases), in such a way that conforming to the narrow convention is automatically also conforming to the wider one. Thus, the conventions that arise from parental commands are meant to also express the community-wide conventions that express the shared principles.[18]

The difficulty in developing nested conventions is that each convention is a regularity in behaviour, and it is hard to see how you could conform to different behavioural regularities at the same time. This is the point that is addressed by my suggestion that the conventions must be related in such a way that conforming to the narrower one is also conforming to the wider. We can identify another kind of underdetermination, one that arises when deciding how to conform to the wider conventions: there may be many different *narrower principles* that are consistent with those principles shared across the whole community. These narrower principles are endorsed by the adherents to the narrower conventions, but not by those who aren't party to the

[16] For an excellent overview of the content of such a multiculturalism within a society, see Michael Walzer, "Comment," in *Multiculturalism*, ed. Amy Gutman (Princeton, NJ: Princeton University Press).
[17] A thought that appears already in Herodotus.
[18] We have already seen some treatment of this when regarding the non-household-regarding parental commands in §IV.

narrower conventions. These can be understood as arising in the process of developing a more articulated view of what their moral behaviour consists in: different people can conclude that different principled ways of acting are what lies behind their conformity to the wider conventions. These narrower principles can then be used either to cover cases that are not subject to wider conventions, or to give more detailed guidance in the same range of cases. Similarly, there can be the accidental or gradual development of ever-more-intricate narrow conventions to the same effect.

To give an example of nested conventions, consider again the example of parents raising their child to be a vegan within a society where all share a devotion to animal welfare, but not all to the extent requiring veganism. The household needs various conventions in order to succeed in being vegan, because it is a strategic situation that requires coordination on their part about things like what kinds of food to acquire and cook in, and not taking on commitments that require them to partake in animal-derived products. So, whereas the vegan households can depend on the wider conventions about the appropriate treatment of animals to avoid the various behaviours which have been settled on as animal cruelty, they also need narrower conventions to avoid contravening their own stricter standards.

There is a disagreement between the vegans and the rest of society about whether the shared principles against animal cruelty necessitate the extra step to veganism. But what there isn't disagreement on are the wider conventions against the various ways to treat animals that have been settled on as instances of animal cruelty.[19] Any stricter standards that the vegans endorse will also automatically meet the standards that are expressed by way of the wider conventions. Thus, we can plausibly model the vegans' commitments as narrower conventions nested inside the wider ones. Just as the community as a whole use the wider conventions to

[19] It bears repeating that of course not all of these settled issues would be settled by convention —for instance, some are likely to be entailed directly by the principles—but the set of these issues will be supplemented by conventions insofar as they are needed to address underdetermination by the principles.

coordinate towards the ends of preventing animal cruelty, the vegans use their narrower conventions to coordinate towards the ends of removing animal products for their lives as best they can. In the same way, we can model other instances where there are multiple different behavioural regularities over the same domain to be found within the same community.

XIII. Conclusion

In this chapter I defended the *commands-as-conventions criterion* for conventional authority, to the effect that it is sufficient (but not necessary) for a command to give those subject to it a pre-emptive reason to conform if doing so would lead to a *limited convention*. A limited convention is a Lewisian convention where outcomes are ranked by the degree to which they conform to the community's shared principles. I introduced the *strategic underdetermination problem* (SUP) as the result of the principles underdetermining what outcome is the best, and individuals in the community need to be able to predict what the others will do in order to know what they themselves should do. I described how the pronouncements of an authority can produce a reliable guide that allow the community to navigate through the SUP case if it meets two requirements: the *salience requirement,* such that there is the general expectation that those subject to the command will conform to it, and the *judgement requirement*, that the authority actually manages to identify an outcome that isn't determinately worse by the lights of the principles than any of the others available. I linked this criterion, which applies command-by-command, to the more common understanding of authority as being vested in an individual or a role by introducing *benign arbiters*, an individual who in every instance satisfies both the salience and judgement requirements, with the result that individuals or roles carry conventional authority to the extent that they approximate a benign arbiter. I then showed how my account counts as an instance of the service conception of theoretical authority, and accordingly slots into one prominent current in the relevant contemporary theory. This manner of approaching the issue has a number of suggestive consequences: it can offer a way of

understanding authoritative commands as ways for the community to conform to their shared principles; it can explain instances of authority that are taken not to be conventional, such as parental authority; and it can perhaps in the future give us a way to make sense of moral variation within communities.

4. The Virtues in Word and Deed

It has recently been claimed by David Velleman that we can't compare moral evaluations across societies because different societies adopt different schemes of action-types. He draws on a wealth of ethnography to make the case that we can't evaluate actions as, say, lying *simpliciter* but only by way of the relevant action-type in the agent's culture. This leads to problems with interpreting actions, because what Anglophone westerners might take to be lying *simpliciter* may in a different culture be a token of an action-type which is an accepted type of not telling the truth, just like Anglophone westerners don't usually disapprove of someone not telling the truth while joking. These include action-types like Russian *vranyo*, Egyptian *kala:m*, Lebanese *kizb*, or Javanese *étok-étok*. These different action-types may all range over the same domain (not telling the truth), but they have different evaluative standards. In his words:

> *The result is that communities can find themselves unable to disagree about what should be ordinarily done, because they differ with respect to what is doable: there is no neutral domain of action-types from which they choose what to do. What's more, action-types are invented, and there is no domain of inventable action-types from which communities can choose which ones to invent, much less disagree about such choices.*[1]

Velleman then goes on to use this claim that there can be faultless disagreements on moral questions between different societies to shore up his case for moral relativism. In a different venue when discussing the same material, he says the following:

> *The problem is not getting the hang of how other communities divide right from wrong, so that we can project from observed cases of their moral judgments to novel cases. The problem is rather that they are dividing a different domain of cases, and so projecting their judgments would still not reveal whether we agree or disagree.*[2]

This is the view I set out to counter.[3]

[1] J. David Velleman, *Foundations for Moral Relativism*, 2nd ed. (Cambridge: OpenBook Publishers). 2.
[2] "Doables," *Philosophical Explorations* 17, no. 1: 12.
[3] In the book-length treatment Velleman also appeals to his views on the perspectivism of reasons and the construction of agency in order to shore up his case for moral relativism. I won't engage with these features of his view, because I'm not trying here to adjudicate on whether we should adopt his relativism. I am concerned

The purpose of this chapter is to attack Velleman's view head-on, by giving a theory of how we can track action-types across societies such that they remain genuinely different from each other but not making substantive engagement impossible in the way Velleman supposes. Here I use the framework of this thesis to indicate how different cultures have limited conventions that specify what the relevant action-types for that culture are. My discussion focuses on the virtues and vices in particular, since they are the most prominent examples of evaluatively loaded action-types, and are widely taken to be the most sophisticated pre-theoretic evaluative framework individuals can draw from.

I will refer to the action-types that correspond to virtues and vices as the *v-types*. Later in this chapter I will expand 'v-type' to also refer to the trait-types that correspond to the virtue- and vice-related action-types. In English these v-types are referred to with terms like 'courageous', 'generous', 'rash', 'greedy', and so on. The v-types are a mix of evaluative and descriptive features, and each highlight ways that these features are meant to characteristically co-occur. For instance, 'courage' ascribes to an agent something like a recognition of the danger in a situation co-occurring with an unwavering pursuit of some worthwhile good that is threatened in this situation.

I adopt unapologetically the view that the v-types can be genuinely different in different societies, even to the extent that some token behaviour is disapproved of in one society by falling under one v-type, and that same behaviour be approved of in another society by falling under a different v-type. By doing this I am accepting the premises Velleman uses to argue for the fact that there can't be substantial evaluations between cultures. In the theory I also present the v-types as not coming from some universal list but each being in some real sense an invention by the culture in question, derived from the substantial and contingent features of

only with countering his view that cross-cultural evaluations aren't possible because of divergences in what action-types different cultures adopt.

those societies. I offer a model for how this process of invention is meant to occur. But on my account the resulting variation in v-types is limited: there are different possible conventions that may be developed in response to the same issue, just like there are different road rules or marriage customs societies may adopt, but all the variations conform to some set of principles.[4]

In the first section of the chapter I elaborate on the puzzle that motivates my account: how to allow both for variation in the v-types as well as their robust cross-cultural similarities. In the second section I outline how limited conventions address that problem: roughly, that the framework within which the v-types occur underdetermines their content, meaning that there are different allowable schemes of v-types available, and the variation in v-types follows from different societies implementing different options. This involves me outlining three descriptive theses that make up the detail of my account. Sections three through six are devoted to these three theses in turn, pausing along the way to highlight their import on debates in the literature on the virtues and vices.

XIV. Principles, stability, and variation in the v-types

It is my contention that we can identify some underlying principles that limit the variation of v-types. The principles I have in mind are the requirements that the variations must all match up to the *evaluative point* of that v-type.[5] The notion of evaluative points has been used in the literature on thick concepts, in that various descriptive features of an action fall or don't fall under a thick concept depending on whether that action matches up to the evaluative point of that thick concept. So, to call something brave is to commend it as a response to danger; to call something mean is to condemn it as holding back something that should be given, and so on. The v-types are obviously thick concepts, as are the localised action-types Velleman appeals to.

[4] This approach shares a lot of common with the influential treatment in Martha C. Nussbaum, "Non-Relative Virtues: An Aristotelian Approach," *Midwest Studies In Philosophy* 13, no. 1.

[5] I take the phrase from Jonathan Dancy, who uses it when discussing 'thick concepts'. Jonathan Dancy, "In Defense of Thick Concepts," *Midwest Studies in Philosophy* 20, no. 1 (1995).

This evaluative point isn't indexed to any particular society or evaluative framework, but on my analysis is instead something that is subject to precisification through conventions. There are various evaluative points to various universal human activities, but what the relevant particular activity is in a given society is something that is sensitive to convention. Using the framework I've developed in the thesis, this is a *limited convention*, because the evaluative point is a feature that precedes and constrains the conventions that precisify it.[6] This means that understanding the adoption of an action-type scheme as a limited convention offers an explanation both of the extent of similarity shared across schemes, and the scope for divergence. Both these features are needed, as I now discuss.

That there is likely to be some underlying similarities can be seen by how often and how naturally we leap across great cultural divides in our appreciation of v-types. For instance, parents read their children Aesop's Fables, despite 2500 years and all of Christendom separating the tale and the audience. Velleman doesn't deny that it is possible to understand each other across cultural boundaries, but he seems to believe this needs the resources of anthropology and ethnography.[7] But in the case of Aesop's Fables, the stories from various mythologies, and countless other examples, we have people in radically different societies taking these action-descriptions to be so straightforward that they are paradigmatically shared with children, and not as an anthropological exercise.

My use of evaluative points as a way to fix cross-cultural reference for v-types contrasts to Nussbaum's approach, where she does the same by pointing to kinds of experiences shared across human lives and the accompanying spheres of action. For my purposes I find the articulation of points and targets to be more useful than articulation in term of spheres of action; correspondingly my language here is less like Nussbaum's and more like the target-

[6] See my treatments of limited conventions in Ch. 1 for more details.
[7] Velleman, *Foundations for Moral Relativism*: 74.

centred conception of the virtues proposed by Christine Swanton.[8] I don't take this to be a deep difference between Nussbaum's approach and mine, and I see us as fellow travellers regarding this issue. This is because to identify the target of an action, such as that the evaluative point of courage is to preserve something of genuine worth in the face of danger, is by the same token to identify a sphere of action, being 'reactions to danger' in the case of courage. So, bridges between Nussbaum's approach and mine are easy to construct.

Strangely, Velleman never tries to address approaches like Nussbaum's or anticipate a response like my own, and takes it as a given that the fact that there are large differences between the action-types recognised across societal borders will hinder cross-cultural understanding. Nonetheless, the many instances of inter-societal understanding isn't the final word on the matter.

If we only needed to give an account of the robust similarities, we could have provided something like a translation schema for the terms for v-types across speech communities. Doing so would be to explain variation in the application of these terms, and correspondingly the variation in the application of the v-types. This would consist in a description of how a v-type in one community maps onto the use of the relevant v-type or -types in another community. This is a familiar kind of task, not straightforward by any means but one which people often accomplish successfully nonetheless. Consider that we have a very large canon of translated literature of great moral complexity. Someone with an interest in such literature is as likely to cite Dostoevsky, Goethe, and ancient Greek tragedy in translation as any text in English. There are notable examples of this approach in the philosophical literature: Rosalind Hursthouse, for one, emphasises the large number of virtue- and vice-terms that have counterparts in every language, and approvingly cites the Virtues Project and their collation of

[8] Swanton, *Virtue Ethics: A Pluralistic View*.

such universally-attested terms.[9] On my account, the similarities that allow us to identify the counterparts of particular v-types in other societies are because their v-types match up to the same evaluative point or points.[10]

But Velleman forcefully makes the point that that mutual understanding sometimes falls away with his many examples of how in different cultures we can come to radically different evaluations of what are substantially similar behaviours. An example of someone else making the same point is when Alisdair MacIntyre notes how we readily recognise as courageous the fatal last stand of the protagonist of an Icelandic saga, but not so for the last stand of the Hitler Youth in the final stages of WWII, despite many pertinent similarities like not turning away from the task despite the obvious danger.[11] So, mere similarity in an evaluative point doesn't secure cross-cultural evaluation.

If we only had to account for the breakdowns in mutual understanding, we could have provided a model where the individual v-types were indexed to some overarching structure which assigned each individual v-type its meaning according to its place in the structure. This is exactly what Velleman proposes. This would preclude the use of a translation schema or similar measure, because then we couldn't make references to v-types across cultural boundaries; we'd instead need to refer in the first instance to the culture-specific structure that v-type features in. But as discussed above, precluding such a translation schema just is to deny the manifest, since we as a matter of course have such translation schemata for sophisticated evaluative material in different societies like Aesop, Dostoevsky, Goethe, etc. Some other account for these differences needs to be offered. The task I set myself here is to preserve the distinctness of v-types in different societies without resorting to the unbridgeable differences in

[9] Rosalind Hursthouse, "The central doctrine of the mean," in *The Blackwell Guide to Aristotle's Nicomachean Ethics*, ed. Richard Kraut (Malden, MA: Blackwell, 2006), 112-14.
[10] On Nussbaum's account this would be because they are identified as right action in the identified spheres of action. I don't take it that there is any competition between my view and hers.
[11] Alasdair C. MacIntyre, *A Short History of Ethics* (Notre Dame, IN: University of Notre Dame Press, 1966). 132.

their use Velleman supposes. I do so by describing how the same evaluative points can receive radically different development in different societies by way of limited conventions.

XV. Virtues and conventions

In broad outlines, my account is that each society has an interrelated complex of v-types that it implements—let us call them *v-complexes*—with there being a range of different workable v-complexes that could be adopted, and it is a matter of convention which v-complex a particular society adopts. The strategic import of such a convention is obvious: all the members of a society have a deep interest in having the same conception of the action-types in play in evaluation, because each of them tailors their actions to their expectations about what the actions in question amount to. This is a point Velleman himself makes clear, which is why Velleman and relativists of a similar stripe argue that the action-types vary between societies but is much more stable within societies. A v-complex is an overarching framework which relates the various v-types to each other, and different v-complexes lead to divergences in the v-types in just the way that Velleman suggests different practices lead to different virtues and vices. But the different v-complexes are meant to be robustly similar to the extent that they answer to some shared underlying framework. The underlying framework in this case—what takes the place of general principles I appeal to elsewhere in the thesis—is the evaluative points of the various v-terms. The v-complex is useful and compelling to its adherents insofar as it makes good on enough of the evaluative points in play.

We can have both similarity and difference between v-complexes if the underlying framework—the evaluative points—underdetermine the content of a v-complex: insofar as this framework is determining it will lead to similarities; where it isn't determinate it allows for differences. The different v-complexes that all match up to the evaluative points would be different options available to populations to adopt, just like Lewisian conventions are called for where there are different behavioural regularities that they could co-ordinate towards. But,

contra Velleman's worry, it isn't that a society picks one of these v-complexes out of something like a society-neutral list; instead there is a process of social construction as the society's conventions make determinate the underspecified content of the framework. To describe this process of construction, I present and defend three theses about the virtues and vices:

1. the conventional fixing thesis:

   *v-types occur within overarching virtue- and vice-complexes (*v-complexes*) which combine them into an overarching evaluative structure, where there are multiple possible v-complexes available, and the prevalence within a community of any particular v-complex and the behaviours that come from subscribing to that v-complex is a limited convention;*

2. the paired profile thesis:

   each v-type is to be identified with a respective pair of profiles—a behavioural profile *consisting of the observable behaviour, and an* intentional profile *regarding the psychology of the individual as they act;*

3. the functional definition thesis:

   *each v-type includes both instances of actions (*v-acts*) and character traits (*v-traits*); v-traits can be defined in terms of v-acts, such that the v-traits are those character traits that in the relevant circumstances lead to their possessor spontaneously performing v-acts.*

I will explain the novel terms that feature in these theses when I come to the sections devoted to them.

The above is one thesis that indicates how this subject matter can be handled by different conventions (1), followed by two others (2 and 3) that deal with the concrete subject matter of the v-types in their own right. It is possible to give an account of conventional variation in the v-types without dealing with the details of their content, but such an account will be necessarily incomplete and I believe would not be very compelling. Velleman has demanded that we give due attention to the fact that localised action-types are the product of a contingent process that he characterises as a form of invention; by describing how we can track the evaluative point of a v-type in all of its permutations as it gets developed within a society I am attempting to make

good on this challenge. Furthermore, in the case of the v-types any descriptively adequate account is going to have to make something like the distinctions between behavioural and intentional profiles and between v-acts and v-traits, and this requires someone who deals with variations in v-types to track how a variation in among one of these leads to variation in the others. It also makes more complete and more compelling my account of how these localised action-types can be invented as different ways to match up to the same range of evaluative points. That is what I attempt to do for the rest of the chapter.[12]

XVI. The conventional fixing thesis

The point of departure for the view I defend here is the *conventional fixing thesis*:

> *v-types occur within overarching virtue- and vice-complexes (*v-complexes*) which combine them into an overarching evaluative structure, where there are multiple possible v-complexes available, and the prevalence within a community of any particular v-complex and the behaviours that come from subscribing to that v-complex is a limited convention*

As an example of v-complexes being fixed by convention I will use a comparison between the standards applied to two different kinds of feudal men-at-arms: Japanese samurai and Italian *condottieri*. It's easy to recognise a samurai's refusal to surrender when fighting for the cause of their patron as courage, but with that comes the realisation of how different the standing and conditions of samurai were: how their place in their society and their status and that of their families was conditional on such extravagant acts of loyalty. In contrast, *condottieri* weren't as wedded to their client's cause, nor were expected to be. Instead, they were independent agents who served clients on a retainer, frequently moving from one employer to another when the terms of their contract were up. The samurai took their standing from being agents of their patron's will, whereas the *condottieri* as a professional class of free agents were judged by the standard of the service they delivered, broadly as lawyers or doctors are. A *condottiere*'s trade

[12] Note that I don't wish to distinguish between 'acts' and 'actions', and use the terms interchangeably.

was battle, which requires courage, there were a range of specific requirements around how to act towards your employer after a contract concludes that amount to standards of loyalty practices, including requirements not to take up arms against former employers for at least two years after the expiry of a contract. Thus, what counts as courage and loyalty for the samurai differs from the same for the *condottieri* by way of the different societal structures they operated in. Nonetheless, courage for samurai and courage for *condottieri* have the same evaluative point: the courageous actions are the appropriate responses to danger when pursuing worthwhile ends. The same goes for the shared evaluative point of loyalty for samurai and loyalty for *condottieri*: they regulate the appropriate relationship between patron and client.

I'm not saying that there is some master action-type 'courage' of which the culture-specific 'courage of samurai' and 'courage of *condottieri*' are species or variants. That would be a view that Velleman strenuously objects to, and I am happy to follow him in rejecting it for present purposes. What I'm saying is that each culture—in this example, feudal Japan, late-medieval Italy, and our contemporary cosmopolitan Anglophone culture—has its own network of v-types coalescing into its v-complex. What allows us to identify some v-type (or set of overlapping v-types) as the one that is closest to contemporary Anglophone 'courage' is whatever the members of those cultures take to be their v-types that match up to the evaluative point of contemporary Anglophone courage (or whichever of its evaluative points is at issue in the particular discussion, in the likely case that one v-type answers to many different evaluative points). It turns out that these v-types are broadly speaking what the feudal Japanese called *yū* and what the medieval Italians called *ardimento*, as can be seen by the fact that nobody has any hesitation in translating references to these v-types as 'courage' in contemporary English.

The translation of 'courage' and other references to v-types across societies isn't meant to be perfect or occur without remainder. In our example, the Middle Japanese word *yū* has notes of austere devotion not always present in the English counterpart, and similarly the Middle

Italian word *ardimento* has notes of daring and audacity not always present in the English counterpart. This is neither a surprise nor a problem. Evaluating the behaviour of men-at-arms involves evaluations of courage, loyalty, and much else besides, and the requirements of one v-type inform the requirements of the other. For instance, if there is some socially settled way to identify the requirements of loyalty, then they also inform the requirements of courage, and *vice versa*. If a certain act of loyalty is socially settled as a goal worth pursuing, then by the same token when its pursuit puts one at risk there will be some response to that risk that would be recognisable as courage. The same goes for instances of the corresponding vices of cowardice or foolhardiness from not appropriately responding to such risks. Some of these interrelations may be one-directional, but many have the influence go both ways: understanding what it means to be loyal specifies objects for our courage, and an understanding of what courage can provide moderates what we think loyalty may ask of us. With a bit of imagination, we can provide many examples of three or four or more such v-types in similar arrangements of mutual dependence: for instance, adding issues of dependability to those of courage and loyalty. Since each of these v-types may itself have further links with yet more of them, there is no point where we can expect a sharp division between what is required for something to count as a token of one v-type rather than another. These interrelations will feature again prominently in §V, when discussing how we use them to define the various features of an action-type.

To gloss what is said above: the v-types have evaluative points, but those points aren't neatly divided up between v-types such that they can be evaluated separately from each other; instead, we need to look at the total system, the v-complex, in order to see how achieving such-and-such an evaluative point is distributed across the v-types constituting that v-complex. Settling one of the constituent v-types is going to have knock-on effects on the others by way of changing the distribution of evaluative points across v-types.

Of course, the circumstances—what the history of the society is, what its neighbours are

like, what kind of climate and environment it finds itself in, what resources are available, etc.—are going to tightly constrain the workable evaluative frameworks for that society.[13] But the above indicates that amongst the determinants of a piece of the evaluative framework are other pieces thereof. It is more than possible that if the pieces of a framework were settled in a different temporal order, then the overall framework would have been different: it is in part because the *condottieri* arose as a class of independent professionals that what loyalty required of them was different from what was required of samurai. Thus, it is possible for societies that are the same before the settling of any such issues to arrive at markedly different societal arrangements, given the contingencies of the process of social settling. This is only one kind of contingency—by the same token, we can expect that at any one stage the scope of evaluations could have been settled differently. These kinds of features of social arrangements are paradigmatically conventional in the everyday (rather than explicitly Lewisian) sense, and that means that we have no reason to expect that either the structure as a whole or any of its individual parts should be uniquely determined.

XVII. The paired profile thesis

In this section I shall set up the descriptive framework I want to use throughout the rest of the thesis. It is the *paired profile thesis*:

> *every v-type is to be identified with a respective pair of profiles—a behavioural profile consisting of the observable behaviour, and an intentional profile regarding the psychology of the individual as they act;*

This means that on my view we describe an action-type as consisting as two related paired

[13] A famous example where a change in the conditions leads to an existing framework becoming unworkable and degenerate is that of the Ik of Eastern Africa, whom first displacement and then famine placed under extreme strain with resultant changes to societal practices (especially in the 1950s-60s). How to best describe and interpret the Ik during that time is controversial, but for my purposes it is sufficient to note examples like individuals trying to escape the surviving practice of reciprocal aid by performing tasks at night so that nobody would see them and offer aid (since such offers would require them to give aid in the future that they might not be able to afford). This indicates an explicit recognition of existing practices and evaluation (that reciprocal aid is normal and required) and the view on the part of individuals that the practice is no longer suitable, hence undermining it.

profiles: we refer to the action-type both as a single action, and also as the trait that has that action as its characteristic expression; both the action and the trait have a behavioural and an intentional profile. I make use of the v-types to make this point, since the split between actions and traits is the most studied example, and making use of such inherently evaluative notions addresses Velleman's concerns about evaluation in terms of culture-specific action-types head-on. In the following section, §IV, I describe how we link these four different profiles (two sets of paired profiles) to each other, and in §V.i I describe how we use the evaluative point of an action-type to track robust similarities in v-types (and action-types more generally) across cultures.

xx. What is a profile?

I characterise acts and traits in terms of profiles of actions. This is meant to be the familiar phenomenon where we identify something by way of its fitting such-and-such profile, or set of profiles. `By 'profile' I mean a collection of descriptions that can be applied to something: the type of identification that we do by way of matching something up to the items on a checklist. Profiles are extensively used in psychology to characterise mental phenomena, and that is the kind of thing I have in mind. One way they are used in psychology is by way of assigning scores (often on a Likert scale) to the behaviour of some subject, with the profile referring then to the sequence of scores assigned to the behaviour across a range of observations.[14] Another use that I have in mind is when we compare some token behaviour to criteria on a checklist. Sometime 'profile' is used in the psychological literature to refer only to sequences of scores, and not to the use of checklists. This isn't a problem, since it is of course trivial to get a qualitative score from a checklist, such as when you say that someone displays 7 out of 8 criteria for A, or meet criteria that scores them above the 90th percentile for B.

[14] See for instance the extended philosophic use of this in Christian Miller, *Character and Moral Psychology* (Oxford: Oxford University Press, 2014). 47-61.

Profiles are descriptions, so they consist of linguistic items: a range of predicates and relations and names, clauses made up of combinations of these. The commonality between a range of token profiles that make them all be profiles of the same type comes from the same predicates being truly predicated to the diverse individuals.[15] Let us take an example from psychology. When discussing, say, insomnia, we find among the diagnostic criteria (after defining 'sleep difficulty') examples like 'the sleep difficulty occurs at least 3 nights per week' and 'the sleep difficulty is present for at least 3 months'.[16] These correspond to the relevant predicates, that then are attached to individuals in order to form the relevant sentences. So, to say that Deng suffers from insomnia is also to say (since these are necessary criteria for the diagnosis) that 'Deng has sleep difficulty at least 3 nights a week' and 'Deng has had the sleep difficulty for at least 3 months'.

The above example is the simplest kind of profile, consisting just of necessary conditions, but a large amount of profiles are clusters of related phenomena. In the psychological case, diagnostic criteria for such cluster cases are often that someone displays, say, five out of the eight listed symptoms. In a cluster case, rather than just checking the truth of all the individual-plus-predicate sentences when matching an individual to a profile, there are further criteria about how the truth of particular individual-plus-predicate sentences corresponds to the full diagnosis. Keeping with our example of sleep disorders, consider the Epworth Sleepiness Scale (ESS) as an example. It is administered through a questionnaire, asking the subject to rate their likelihood of dozing in eight specified situations on a scale from 0 (would never doze) to 3 (high chance of dozing). These numbers are then added up and result in the subject's ESS score, with anything higher than 12 normally being taken as a sign of 'excessive daytime

[15] 'Truly predicated' means that the propositions that result from applying these predicates to those individuals are true.

[16] "Sleep-Wake Disorders," in *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association).

sleepiness' and that medical intervention is called for. So, the ESS specifies a profile for excessive daytime sleepiness, which consists of 24 predicates corresponding to the situations, one set of four possible scores for each of the eight specified situations—e.g. 'would never doze while sitting and reading (scored 0)', 'has a slight chance of dozing while sitting and reading (scored 1)', etc.—and a further predicate about the resulting score—e.g. 'has a total score of over 12 and thus has excessive daytime sleepiness'—meaning that the type-profile consists of 25 predicates. So, to say that Deng has excessive daytime sleepiness means that the type-profile's predicates are truly predicated of Deng such that the 25th predicate about Deng's resulting score is truly predicated as well. A large range of different sets of these predicates being truly predicated to Deng would suffice for this: if he scored 3 on four situations, 2 on six, or any other set of scores that would lead to a total score of 12 or more. As for the ESS, so too for any other profile that comes in degrees, as is likely for at least some number of v-types.

xxi. Profiles of v-types

To give example profiles of v-types, I'll continue my use of the example of courage among men-at-arms. The behavioural profile of *yū* (courage among samurai) includes not surrendering in battle, seeking out your peers on the opposing side and challenging them, and avoiding dishonourable acts even at the cost of your own life. Its intentional profile traditionally is meant to include valuing your own life lightly, maintaining clarity of mind and purpose, and not being motivated by fear or acquisitiveness or vainglory. The behavioural profile of courage among *condottieri* (*ardimento*) includes staying steadfast in battle, not hesitating to undertake feats of daring, but also to avoid feats of bravado that undermine their side's chances in battle. The intentional profile traditionally is meant to include taking delight in feats of arms, a desire to engage in exploits, and a clarity of purpose. In §V I provide a theory of how we can make principled and informative links between these behavioural and intentional profiles, but for the time being it is enough to content ourselves with the reports of contemporaries.

Splitting the characterisation of action-types into two profiles is meant to latch on to an important split in how actions are presented to us. There is the bare behavioural presentation, which is relatively straightforward to understand: the movement of a person's limbs, the words they say, their facial expressions, and so on. There are also the psychological factors lying behind an action: the agent's motives, their understanding of the situation they are reacting to, the intended effect of the action, and so on. Accordingly, the action can be presented as the result of such-and-such psychological factors. To settle on a definition, I call the behavioural profile the (in principle) observable movements of an individual in a situation as they perform the action. I call the intentional profile the set of motivations, conceptions, perceptions, and similar psychological features in play in producing the action.[17]

Some may object that we can often directly perceive someone's intentions. For instance, it's certainly hard to avoid describing certain behaviours as being instances of anticipation, but 'anticipation' is a propositional attitude with its object some not-yet-actual outcome. So, if a description is of someone anticipating something, then that is an intentional feature as a I categorise them, despite the fact that often 'anticipation' is meant to be directly observable.[18] I am sympathetic to this view, but for present purposes I wish to work on the terms of someone who wishes to prioritise purely behavioural features, and show in §V that nonetheless we end up with intentional features alongside the behavioural.

We must note that there is no one-to-one correspondence between behavioural and

[17] A note on the name—I don't mean that the intentional profile only includes intentions in the sense of plans or purposes. I mean 'intention' in the old sense Kant uses the term, as anything that is presented to the mind. Intentions in the narrower sense is just one species of this genus, and it turns out that this narrower sense is the one that we are most often concerned with. It is of course commonplace that a genus becomes typified by the most prominent species that falls under it. I could have called this the 'psychological profile', but this could have been confusing because psychologists themselves very often characterise actions in terms of their behavioural profile, and I take most of what I say about behavioural profiles from psychology.

[18] The idea that intentional features are often directly observed in action arises both in work following Wittgenstein, and also in work following from phenomenology. For an example of each in turn, see Rosalind Hursthouse, "Intention," *Royal Institute of Philosophy Supplement* 46(2000); Shaun Gallagher, "Direct perception in the intersubjective context," *Consciousness and Cognition* 17, no. 2 (2008).

intentional profiles. Instead, some cluster of related behaviours is the possible product of a cluster of related intentions, and *vice versa*. There are many different types of behaviours that can be produced by a particular set of intentions: for instance, the realisation that you have been cheated can lead to you displaying anger behaviour, or resignation behaviour, and so on. Similarly, there are many different intentional profiles that may match up to any one behavioural profile: someone fidgeting when being questioned may be because they have something to hide, or they are averse to being confronted by an authority, or they have ulterior motives relating to their interlocutor but unrelated to the current line of questioning, and so on.[19]

The paired profile thesis is meant to be general across instances of action evaluation. For instance, we may usefully characterise consequentialism as the moral theory which says that only the behavioural profiles of actions count for action evaluation, since the consequences of an action are part of its (in principle) observable manifestation in a situation.[20] For instance, elsewhere I use it to analyse how we can conform to something without explicit knowledge of what we conform to (something which is a common claim about rule-following).[21] But it is especially salient for virtue theory since it latches onto a distinction in Aristotle between *acting from virtue* and *acting according to virtue*. To act according to virtue (or vice) is to reach the end that a virtue specifies for the relevant type of action: regarding honesty, to tell the truth; or regarding the vice of meanness, to show undue reluctance to part with one's money even if it's called for. To act from virtue is to act according to the virtue from your own possession of the

[19] We normally describe behaviours only to some degree of detail, and normally that degree isn't very high. Even in the behavioural sciences, where you may have expected the most detail, broad strokes are commonplace: for many social psychological experiments the relevant degree of detail is nothing more than e.g. whether someone helped pick up dropped papers or not.

[20] This highlights the distinctive feature of Julia Driver's consequentialist virtue theory: it makes the psychology of the agent to be a non-factor. Julia Driver, *Uneasy Virtue* (Cambridge: Cambridge University Press, 2001). This feature makes it unsuitable as a virtue theory, as argued by Nancy E. Snow, *Virtue as Social Intelligence* (New York, NY: Routledge, 2010). 5-6, 10.

[21] This is discussed further in Chapters 5 and 6.

virtues, and *mutatis mutandis* for the vices.[22] To illustrate, it is possible and commonplace to succeed at some activity only with the help of some prompt, in contrast to doing it from your own devices: like a chess player who makes a series of game-winning moves after being told what to play by a grandmaster, as compared to a player who makes the same moves because they themselves saw that those are the moves to play. Succeeding at an action only by being prompted means that the agent displays only the behavioural profile of an action, whereas if they perform the action under their own self-control they display both the intentional and behavioural profiles of an action.[23]

It is worthwhile to compare my approach to one adopted for philosophic purposes by Nancy Snow[24] and Christina Bicchierri[25] (independently of each other). This approach is to categorise an individual's behaviours both by an *objective* and *subjective construal*, where the objective construal is how the act is categorised making use of publically observable features of the agent in their situation, while the subjective construal is how the agent themselves categorises the action.[26] It is clear that the behavioural profile of an agent and the objective construal of their action is the same thing. But the intentional profile of an action contains the agent's subjective construal of the situation, and extends past it. An agent's subjective construal of a situation can change as a result of various features that are themselves part of the action's

[22] I follow Christine Swanton in my treatment of this distinction: Christine Swanton, "A Virtue Ethical Account of Right Action," *Ethics* 112, no. 1 (2001).

[23] Insofar as doing something under your own self-control is displaying the intentional profile of that action, which in turn leads to displaying the behavioural profile. This qualification is to circumvent genuine but uninteresting counterexamples: someone has the right intentional profile, but there is some other feature of the situation which leads to the agent accidentally also displaying the right behavioural profile. For instance, somebody is hypnotised in order to keep their arms by their side, except to raise one arm when the phrase 'Ruby Tuesday' is uttered, and ends up accidentally doing so appropriate to a situation when someone asks "who here likes the Rolling Stones song 'Ruby Tuesday'?" Our subject's arm rises because of the hypnosis, but since they like the song and intend to raise their arm to indicate that they like the song, they display both the intentional and behavioural profiles for the act 'raising my arm to show I like the song'; despite this, it isn't the intentional profile that leads to them displaying the behavioural profile.

[24] Snow, *Virtue as Social Intelligence*.

[25] Bicchieri, *The Grammar of Society*.

[26] The objective construal is sometimes also called the 'nominal' construal.

intentional profile: as their attention gets drawn to various features of a situation, depending on whether they have a sensitivity towards such-and-such features, and so on. The import of these extra features can be appreciated by looking at sophisticated references to such features in our evaluative vocabulary. Consider, for instance, terms like 'careless', 'conscientious', and so on. These are evaluations of an agent's attention: whether they give the right amount of attention to the (not yet specified) important features of a situation. When we describe someone as conscientious or not we haven't as yet said anything about the content of their subjective construal, but instead have commented on the manner in which they form their construals. For instance, it is also possible to note that someone's construal may be correct (or laudable, etc.) but that they were careless nonetheless, such as when their construal is correct only by accident. We also have v-types that concern how one person relates to another's intentional profile. Being humourless is an example, at least when used in interpersonal contexts, e.g. when someone is called humourless because they fail to appropriately moderate their reaction to an embarrassing situation in recognition that it was meant as a joke.[27] Examples of such evaluations show that the intentional profiles of actions form a substantial part of our evaluative practices in a way not exhausted by subjective construals.

XVIII. The functional definition thesis

It is clear that our evaluations in terms of v-types are sometimes applied to actions, and sometimes applied to character traits. The fact that these are two different kinds of evaluations becomes pertinent when we consider evaluations of some action (virtuous or vicious) as being out of character for someone: we both lament that generally honest people can be driven to lie by some circumstances, and herald the fact that sometimes someone can surprise others and themselves with conspicuous acts of bravery. The contrast that these observations depend on

[27] There are also non-evaluative uses of 'humourless', such as 'she is humourless about racism', meaning that she doesn't think racist topics are light enough to joke about. These don't enter our discussion here.

just is the contrast between using a v-type to evaluate a trait and using a v-type to evaluate an action.[28] My next step is to indicate how we can go from descriptions of actions to descriptions of character traits. This is obviously of interest for the action-types I'm discussing here, those concerning the virtues and vices, since those are paradigmatically expressed by way of settled dispositions of character.[29] I will in what follows restrict myself to just talking about v-types rather than action-types in general. But there is no reason to suppose that the kind of link between acts and traits will be restricted to v-types, since there are many other action types which are socially settled in the kind of way of interest here (such as those attached to particular social roles). We can illustrate the general interest of what follows by showing why Velleman must adopt the same relativism of trait-types as he does of action-types if he is to make his theory have bite.

It is a dictum that to have a certain character trait means that one will at least typically act in specific way—produce token behaviours that fall under some action-type. While this isn't discussed by Velleman, this link between traits and actions means that for his theory to have the kind of broad import he wants it must not just be action-types that vary from culture to culture, but also trait-types. This is because Velleman strenuously denies that there is anything like a neutral domain of possible action-types from which different societies draw.[30] But if it is only the action-types that are meant to differ and not trait-types, then there will in fact be such a domain: the actions that are instantiations of the cross-cultural scheme of traits. This would severely attenuate Velleman's relativism: there will be this cross-cultural domain of action-types that correspond to the cross-cultural domain of trait-types, meaning that there will be

[28] Thomas Hurka makes a distinction like mine, taken up by Gopal Sreenivasan calling the reference to v-types as applied to actions *local uses*, and their reference as applied to character traits *global uses*. They then go on to defend the view that local uses are primary. I counter their view about the primacy of acts in Appendix A. I don't follow this usage, since 'local' and 'global' are overloaded terms, even within the context of this chapter. Thomas Hurka, "Virtuous acts, virtuous dispositions," *Analysis* 66, no. 1 (2006); Gopal Sreenivasan, "Disunity of virtue," *Journal of Ethics* 13, no. 2-2 (2009).

[29] In what follows I use 'character trait', 'settled disposition of character' interchangeably.

[30] Velleman, *Foundations for Moral Relativism*: 55.

straightforward cross-cultural comparisons and evaluations of actions that fall under these action-types. There would still be room for radical differences in the action-types between different cultures, but this would be in the margins around a solid core of robust similarities. This would leave us with a somewhat anodyne version of cross-cultural differences, one that is often articulated in order to disarm the worry that cultural variation makes universal ethics impossible.[31] But Velleman doesn't want to disarm the worry, so he needs to deny that there is a neutral domain of trait-types.

The above settles that it isn't enough of us just to talk about action-types, but we also need trait-types. In keeping with the rest of the chapter, I will concentrate on the virtues and vices when discussing these. I will call the action-types related to virtue and vice *v-acts*, and the respective trait-types *v-traits*. 'V-type' thus now refers to the genus of which v-acts and v-traits are species.

Since I wish to endorse with Velleman the cultural diversity of action-types, I also endorse the cultural diversity of trait-types, and in this section I offer my analysis of how these trait-types are (in Velleman's terms) invented by different societies. *Contra* Velleman, I also hold that this diversity of action-types doesn't mean that we can't meaningfully evaluate actions cross-culturally. For this section the matter of how we track related action- and trait-types cross-culturally doesn't enter the picture, but I discuss it in §VI. Here I concern myself only with linking actions to traits, which I do by way of what I have called the *functional definition thesis*:

> *each v-type includes both instances of actions (*v-acts*) and character traits (*v-traits*); v-traits can be defined in terms of v-acts, such that the v-traits are those character traits that in the relevant circumstances lead to their possessor spontaneously performing v-acts.*

While in the commonsense understanding we don't find a separation of actions from traits or

[31] E.g. in the treatment relativism receives in the introductory textbook Rachels and Rachels, *Elements of Moral Philosophy*: 15-30.. See also the discussion of the same in the Introduction.

behaviours from intentions, but all of them together in a way that isn't trivial to separate. But very many theorists want to prioritise behavioural features, often because they are straightforwardly observable in a way these theorists doubt intentional features are meant not to be.[32] I will take these two schools of doubt on their own terms and start off only with behavioural profiles of acts, what is meant to be the most secure kind of description available. Then I introduce machinery (commonplace in the philosophies of science and of mind) to show how we can go from these behavioural profiles of actions to intentional profiles and specifications of traits. With this I hope to vindicate the commonsense understanding of all these features, behavioural and intentional, relating to acts and traits, as intimately interlinked. I take the view I offer here to be a precisification of this commonsense understanding, but for present purposes it doesn't matter whether I am right about that.

xxii. Introducing functional definitions

The reason I introduce functional definitions is to make use of the so-called Ramsey-Lewis method to forge a bridge from acts to traits.[33] This is meant as an answer to an objection in the virtues literature against views like my own. The objection is made by Gopal Sreenivasan against what he calls 'post hoc virtues'. He grants that giving a definition of v-traits by way of v-acts (using my terms) is a plausible way to analyse references to v-traits, but objects against their use on the grounds that they give specifications of the virtues that lack independent content. This is because he takes it that they can only be a restatement of what the relevant acts are, and don't give us further grounds on which we can identify a particular trait as virtuous or

[32] Similarly, there is a trend in recent philosophy and psychology that casts doubts on whether there are traits of this kind at all, or (in weaker versions) whether the traits deserve to stand on the same footing as acts. I discuss this skepticism, and how the functional definition thesis responds to it, in Appendix A.

[33] This is equivalent to approaches that handle functional definitions by way of machine state diagrams. It is possible to go from one approach to the other, as Mark van Roojen does in his treatment of what he calls 'network analyses'. Mark van Roojen, *Metaethics: A Contemporary Introduction* (Oxford: Routledge). 237-52.

not.[34] We can render Sreenivasan's objection in the form of the following argument:

1) *Take some particular set of behaviours as given.*
2) *We can give a post-hoc specification of the respective trait: the trait is whatever settled psychological features of the individual typically lead to them performing the given behaviours.*
3) *The post-hoc specifications of traits contain only the same information as the given behaviours.*
4) *Any worthwhile specification of traits will be more informative than the given set of behaviours.*
5) *Therefore, the post-hoc specification of a trait from some given set of actions isn't worthwhile.*

However, the consensus in the philosophies of mind and science has concluded that objections like these are too pessimistic. It is now a commonplace that functional definitions by way of Ramsey sentences can provide definitions in terms of functional roles (for mental states, and generally) with independent content. This is in explicit recognition of the fact that they are post hoc in just the way Sreenivasan objects to. In the above argument, this claim amounts to a denial of (3), to the effect that we can offer a post-hoc specification of a trait that is more informative than just the behaviours we use as inputs. In particular, the traits garner their content not just from the given behaviours, but also the relationships that hold between the post-hoc traits. The Ramsey-Lewis method does this in general for any post-hoc specification of theoretical terms from some given information, and this is standardly done where the theoretical terms are mental states or traits in particular [35] This means my approach escapes Sreenivasan's criticism.

One important feature of the Ramsey-Lewis method is that when using it we don't specify

[34] In his words: "To explain its possessor's behaviour, a disposition has (minimally) to be describable in independent terms, i.e. independent of the behaviour (or output) in which it results." Sreenivasan, "Disunity of virtue," 210-11.

[35] Jackson in an introductory textbook co-written with David Braddon-Mitchell discusses this point directly: David Braddon-Mitchell and Frank Jackson, *The Philosophy of Mind and Cognition*, 2nd ed. (Oxford: Blackwell). 55-59. The same point is made by Stephen Yablo, "Definitions, consistent and inconsistent," *Philosophical Studies* 72, no. 2-3.

definitions for theoretical terms one-by-one, but instead do so for all the terms used in a particular theory at once. This is because we use the theory to generate a network of functional relationships, and the various theoretical terms are meant to be the occupiers of these roles. If we change any one of the terms, we change the whole network, which requires defining all of the terms all at once.

The Ramsey-Lewis method is an approach for defining theoretical terms, usually called the *T-terms*. T-terms are contrasted to the terms we don't need a theory to understand, usually called the *O-terms*. When Carnap introduced the term he meant these to be 'observational terms', but Lewis wants them just to be 'old terms', ones for which we have established agreed-upon interpretations. Lewis has the O-terms each be a name: a name of an individual, a relation, a predicate, and so on. So, both the O-predicate 'wears size 10 shoes' and the O-relation ' ___ has been exposed to ___' are examples, as is the copula and other similar parts of our everyday vocabulary. All the names used in the theory we don't have established agreed-upon interpretations for, the ones introduced by the theory, are the T-terms. They are the ones we want a theory to provide us an interpretation for. In this case, since we are granting critics like Sreenivasan that bare behavioural descriptions of one-off actions are to be taken as primary, we accept only behavioural profiles as O-terms and leave the intentional profiles and the v-traits as T-terms, to be settled with this method.

Let us call the statement of a particular theory its *postulate*. The usual way to do this is to make a sentence consisting of the conjunction of all the individual statements that are part of a theory. It is often useful to have a label for a postulate where we note in brackets what the terms feature in the postulate, normally the T-terms. So, a postulate of theory T with T-terms t_1, t_2, t_3, …, t_n would have the label '$P_T(t_1, t_2, t_3, \ldots, t_n)$'.

We can read the postulate as an implicit definition of the T-terms: the T-terms refer to whatever it is that makes the postulate true. We can make this implicit definition explicit by

making use of *Ramsey sentences*. The Ramsey sentence of a theory is the statement that there is a set of individuals that count as *realisors* for the theory, meaning that all the various claims made of the T-terms are true of the realisors. You create the Ramsey sentence of a theory by replacing all occurrences of the T-terms in a postulate with the use of variables (higher-order variables, since they also refer to predicates and functors), and existentially quantifying over the postulate for the existence of entities that satisfy those variables. So, the Ramsey sentence R_T of the theory with postulate P_T is:

$$\exists x_1, x_2, x_3, \dots, x_n . P_T(x_1, x_2, x_3, \dots, x_n)$$

Now that we replace all the T-terms with variables, the postulate contains no T-terms. The variables are just placeholders, and all the content-bearing terms are O-terms with settled interpretations. But what has remained after the T-terms are replaced are the various relationships between all the terms in the theory. This is something that isn't contained just within the O-terms, but comes out in the full statement of the theory. In this way applying the Ramsey-Lewis method on a network of relations between O-terms in the postulate of a theory provide us with informative and explicit definitions of the theoretical terms: they are the individuals that match up to all the clauses in the postulate that involve the variable in question. This is expressed by way of identifying a particular T-term with a particular member of the set of realisors: e.g. 'T-term 1 is identified with the first member of the set of realisors, T-term 2 with the second member, etc.'.

Lewis gives a nice toy example of this approach. Consider a murder mystery, where a detective gathers all the parties to the case into a room and starts laying out the facts. The detective introduces the labels X, Y, and Z, to track the as-yet-unidentified individuals responsible for the murder. So, the facts of the case in the detective's telling include 'X, Y, and Z conspired to murder Mr Body', 'X was Mr Body's partner in the gold fields of Uganda', 'Y and Z conferred in a bar in Reading last week', and so on. The facts of the case amount to a

theory about the murder of Mr Body, and the aggregate of all of these facts form the postulate of that theory. We can read this postulate in two different ways. 'X', 'Y', and 'Z' may be names, meaning that they are theoretical terms because we don't have a settled interpretation of these names—we aren't told who X, Y, and Z are. Or alternatively, we can read X, Y, and Z as variables in a Ramsey sentence, meaning that we don't look to them for information but instead use them as ways to relate the various individual facts to each other. If we interpret the detective's account as a Ramsey sentence, we find an explicit way to identify the murderers: there is a triple of individuals X, Y, and Z, such that the facts of the case are true of those three, and X is the person who is the first realisor of the set, Y the second, Z the third. So, if we find out that all the facts of the case are true of Plum, Mustard, and Peacock, with Plum being Mr Body's partner in the gold fields of Uganda, Y and Z had conferred in a bar in Reading last week, etc., then we know X is Plum, Y is Mustard, and Z is Peacock.[36]

If there was a mistake in the theory such that nobody matched the given descriptions, the T-terms wouldn't have any definite meaning; the same goes if the theory is underspecified and didn't pick out any three individuals in particular. Lewis addresses these worries by providing analyses for the case of multiple realisors and what he calls 'near-realisors', where nothing realises the given theory but there is a realisor for a slightly modified version of the theory.[37]

xxiii. Functional definitions for paired profiles

My proposal is to have a sequence of functional definitions linking first behavioural profiles of a v-act to their intentional profiles, and then the intentional profile of a v-act to that of a v-trait. The terms that come out at each step of this process feature in our everyday evaluations concerning the virtues and vices: the target behaviours we start with, in the middle the

[36] David Lewis, "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50, no. 3: 250-52.
[37] Ibid., 252-53.

psychological features of actions that I've discussed in §IV.ii, and at the end the settled dispositions of character. We could have left the intentional profiles of v-acts out and gone straight from behaviours to traits in one step, but I think it is clearer if we stagger the process.

We first functionally define the intentional profile of a v-act as whatever collection of psychological features turn out to realise the Ramsey sentence that takes the behavioural profile of that v-act as the O-terms. Then we repeat the process, but this time holding the intentional profiles of individual acts as the O-terms, and positing (as our everyday notions do) that there are dispositions that lead to the individual actions that are seen to be manifestations of them. Whatever psychological dispositions turn out to fulfil this second functional definition are the v-traits. This time, the intentional profile of the v-act counts among the O-terms, and the resulting theory gives a functional definition of the intentional profile of the trait. There are two functional relationships here: between having such-and-such an intentional profile and displaying a corresponding behavioural profile, and between having the trait that leads you to characteristically display such-and-such intentional profile. And when those traits and acts are the ones we note as vicious or virtuous, then we have the v-traits and v-acts.

Before I get to applying the Ramsey-Lewis method to v-acts and v-traits, I'll illustrate how it is meant to go by way of a toy example. Consider being good at pinball. In everyday evaluations of someone's pinball skill there is a mix of behavioural and intentional features, such that someone gets high scores, that they know when to tilt the table, etc. But a critic may worry that only the behavioural descriptions have a firm grounding, and to talk about intentions or traits isn't as certain. No matter, because we can take the behaviours that are taken to be indicative of good pinball play and give functional definitions of the relevant intentions and traits from that starting point.

Here is a theory about good pinball play:

> *A good pinball player rarely misses the ball by timing the strokes of the flipper so they hit the ball, and by being able to aim the ball, moving it away from areas the player*

anticipates will make the ball get to the out hole. A good pinball player gets high scores by having a scoring strategy and having the aim to hit the bumpers and targets in the right sequence to follow that strategy.

This theory as a whole is a definition of 'good pinball play', but in order to cash it out fully we need to explain the other theoretical posits it contains. If we restrict ourselves only to describing people's behaviours, the O-terms for being good at pinball include 'gets high scores', 'rarely misses the ball', as well as the various features of the pinball table like the ball, the flippers, the various bumpers with their respective scores, and the out hole where the ball goes if it misses the flippers. That means the T-terms are those that involve the player's intentional features, which can't be directly observed. In the toy theory above, they are: 'aiming', 'timing', 'anticipation', and 'scoring strategy'. These T-terms are the building blocks of the intentional profile of the action-type 'good pinball play'.

We use the Ramsey-Lewis method to derive the relevant functional definitions—take the postulate given above, replace the occurrences of these T-terms with variables, and posit the existence of realisors for the theory, identifying the T-terms with the members of this set of realisors—resulting in the following Ramsey sentence (in natural language):

There is an x_1, x_2, x_3, and x_4, such that a good pinball player rarely misses the ball by means of using x_1 to match the strokes of the flippers so they hit the ball, and by means of x_2 moving it away from areas the player has the x_3 that it will make the ball go into the out-hole. A good pinball player gets high scores by having a x_4 and by way of x_1 to hit the bumpers and targets in the right sequence for x_4.

Thus, 'timing' is functionally defined as whatever x_1 turns out to be, being whatever works along with x_2, x_3, and x_4, such that it makes the player make the flippers move such that they hit the ball, etc. There will be many layers of description available of these realisors. For most of us most of the time we're just interesting in the brute success of someone managing to hit the ball with the flippers when they intend, and only care that there is some realisor for this functional role rather than caring about what exactly it is. But there will be something that counts as the realisor of the functional role 'timing'. On the physicalist story, for instance, this

will be the patterns of neuron firing such that they coordinate the movements of the fingers such that the ball will move to such-and-such a point in their visual field. This allows us to cash out our theoretical terms at every level of description, from the breezy high-level descriptions we have available merely by our competence in a language, to the painstaking detail that concerns neuroscience (in case the physicalist story is true). And so on, for the other T-terms. And in this way we have the intentional profile of the action-type 'good pinball play'.

Now we introduce a further theory, one about the trait of being good at pinball:

> *Being good at pinball involves good reflexes, to help with timing and not missing the ball when it moves quickly and sharply changes direction, and the focus that keeps their aim from slipping or losing track of their scoring strategy.*

Now that the above functional definition for the intentional profile has been put in place, we have settled interpretations for the terms 'aiming', 'timing', and 'scoring strategy', and can use them without concern. This means the T-terms for our theory of the trait of being good at pinball are 'reflexes' and 'focus', which are traits that the theory posits as what allows the reliable and spontaneous production of the intentional profile of the action-type good pinball play. Again, using the Ramsey-Lewis method we swap the T-terms out for variables in the postulate, make an existential quantification over the result, and then produce functional definitions by defining the T-terms as the respective realisors of this Ramsey sentence. So, 'good reflexes' will turn out to be whatever turns out to keep someone having good timing even involving some fast and erratically moving—if our current neuroscience is correct, various arrangements of the cerebral cortex and the condition of fast-twitch fibres in the finger-muscles of the pinball player. In this way we started with the bare behavioural profile of an action-type, and derived first the respective intentional profiles and then the respective trait-type, all of which having different realisors but being parts of the same functional system.

xxiv. Functional definitions for v-types

Let us now move to the real subject matter, v-acts and v-traits. I'll take *yū* (courage for

samurai) as our example. In §VI I'll compare it to *ardimento* (courage for *condottiere*) in order to highlight how conventional differences flow on between the various parts of a v-complex. Let me repeat the statement of the behavioural profile of *yū* from §IV.ii: it includes not surrendering in battle, seeking out your peers on the opposing side and challenging them, and avoiding dishonourable acts even at the cost of your own life, etc. Let us now do the same for the theoretical terms we're introducing, the constituents of the intentional profile, but this time decomposing it into individual statements, which I will use as the T-terms of the functional definition. On the traditional understanding, the intentional profile of *yū* involves a range of intentional features, some of it having names of their own, some of it unnamed.[38] To not strain my or the reader's command of medieval Japanese, I'll give all these intentional features labels, as well as a natural language description in contemporary English that is meant to denote them. Properly speaking these labels and the corresponding words in medieval Japanese are the theoretical terms, not the descriptions I give them, which are at most an incomplete grasp of those features. We depend on a theory to flesh out these incomplete descriptions into definite and informative specifications:

> t_{Y1}—*valuing honour,*
> t_{Y2}—*keeping your wits about you,*
> t_{Y3}—*aiming for a clarity of purpose,*
> *etc.*

The postulate of *yū*, let us call it P_Y, is a description of some range of behaviours as following from such-and-such intentional features. That is, it's a conjunction of clauses that involve both O-terms (here, the constituents of the behavioural profile) and T-terms (the constituents of the intentional profile). The O-terms aren't exhausted by the behavioural profile of *yū* but range

[38] In a different context Aristotle already notes that it is possible that there won't be a common name for every item that features in our full explanation of virtue and vice as it occurs in everyday evaluations, when he notes that his contemporaries don't have a common name for the vice of excess relative to temperance. Aristotle, "Nicomachean Ethics," 1118b-19a.

over all kinds of things that we don't need the introduction of intentional profiles to explain, like facts about what battles were like in feudal Japan. Let us give a following postulate for the traditional view of *yū*, $P_Y(t_{Y1}, t_{Y2}, t_{Y3}, \dots)$:

> *The samurai should not surrender in battle because they value honour and surrender is dishonourable; they should seek out and challenge their peers on the opposing side because to do so is the most effective way to win the battle and their clarity of purpose demands they do so, despite the fact that this is the most dangerous way to fight, to turn away from the demands of honour through fear would be to fail to keep their wits about them; etc.*

We take this postulate, and replace all the mentions of the T-terms with higher-order variables $x_1, x_2, x_3,$ and so on for the other t-terms. Then we perform an existential quantification over those variables to arrive at the Ramsey sentence of the theory, R_Y:

$$\exists x_1,x_2,x_3, \dots, x_n.P_Y(x_1,x_2,x_3, \dots, x_n)$$

We can write R_Y out in natural language as:

> *There is an $x_1, x_2, x_3,$ and so on, such that the samurai should not surrender in battle because surrendering would be inconsistent with x_1; they should seek out and challenge their peers on the opposing side because to do so is the most effective way to win the battle and their x_2 demands they do so, despite the fact that this is the most dangerous way to fight, which doesn't faze them because of x_1 even at the cost of their own life; to turn away from the demands of honour through fear would be to fail at x_3; etc.*

We then get to definitions of the theoretical terms by lifting out the clauses that refer to the respective variable:

> 't_{Y1}' *is the first member of the set* $(x_1,x_2,x_3, \dots, x_n)$*, being the feature that makes you not surrender in battle even at the cost of your own life; that makes you not be fazed by challenging your peers on the opposing side despite the fact that this is the most dangerous way to fight; etc. ; and the set as a whole realises the theory.*
>
> 't_{Y2}' *is the second member of the set* $(x_1,x_2,x_3, \dots, x_n)$*, being the feature that makes you seek out and challenge their peers on the opposing side because to do so is the most effective way to win the battle; etc.; and the set as a whole realises the theory.*
>
> 't_{Y3}' *is the third member of the set* $(x_1,x_2,x_3, \dots, x_n)$*, being the feature that makes you not turn away from the demands of honour through fear, even if it means you accept death rather than act dishonourably, and this doesn't faze you because of x_1; etc.; and the set as a whole realises the theory.*

If the traditional story is correct there will be features like this, and whatever other features are

appealed to in the full theory of *yū*, and they will be broadly the kinds of things we describe as valuing honour more than your life, etc. They will be psychological features and be the kind of things that psychological features turn out to be: patterns of activation of neurons, excitations of the soul, features of the form in the hylomorphism of an individual, or whatever. The theory of *yū* that underpins the evaluation of a samurai's response to danger succeeds if there is such a realisation, and is a failed posit if there isn't, or if we take into consideration that we are talking about a varied tradition rather than a single definitive statement of the theory, some precisification of the theory of *yū* has a realisation. And as for *yū*, so for any v-type, and any action- or trait-type in general.

For the final step we repeat the same process to arrive at the intentional profile of the trait. We do so after we incorporate the intentional profile of the v-act into the O-terms, since we now have a settled interpretation for them by way of the definitions derived in the manner discussed above. The traditional theory of a v-trait like *yū* is going to cite its role in the production of certain behaviours and especially certain intentional profiles of individual actions. For the most part this just will be examples of the disposition-instantiation relationship, with the trait 'disposed to keep your wits about you' producing instances where you keep your wits about you. But just as there will be many links between distinct v-types at the level of actions (courage bleeding into loyalty bleeding into reliability, etc.), so too at the trait level. For instance, many contemporary discussions of *yū* highlight the extent to which a samurai's response to danger involves a certain austerity of spirit. This isn't quite the same as not valuing your own life or having a clarity of purpose, but is non-accidentally related to them. The link is supposed to be that by cultivating austerity as a character trait the samurai is less likely to form attachments to worldly items which can act as distractions, up to and including an attachment to their own life. This encourages not only their willingness to risk their life, but by way of clearing their mind helps them with maintaining clarity of purpose and keeping their wits about

them.

xxv. Going from intentions to actions

The above shows how we can arrive from the v-acts to the v-traits by way of a chain of functional definitions. We may as well have started from the other end and looked at the dispositions to have certain motives, perceptions, etc., and seen what behaviours typically follow. In the above I haven't appealed to the behavioural profile of the trait, but it is easy to specify: it is the behaviours you are disposed to instantiate if you have the kinds of intentional profiles that the trait characteristically produces. Thus, the behavioural profile of the v-trait should just be to characteristically display the behavioural profile of the respective v-act in the relevant circumstances. While this has played no role in the functional definitions, it allows us to close the circle, in a manner of speaking, and is relevant if we take intentions or traits as our starting point, rather than behaviours.

There is a rich tradition of ethical theory that gives pride of place to the motives of agents, including virtue theories. There are problems with making motives and other psychological factors be the first movers of an ethical theory, especially regarding definite action evaluation, since accurate interpretations of behaviours are easy to come by but equally good interpretations of intentions are much harder won. Nonetheless, some theorists have attempted to get at the v-types in this manner—Michael Slote's motive theory of virtue ethics, where the virtuous acts are those that come from the right motives (with no other quality necessary),[39] or Linda Zagzebski's divine motivation theory, where the virtues are the exemplar acts that arise from the motivations distinctively pursued by the exemplar agent.[40]

This difference of direction is easily accommodated. We don't need to appeal to something

[39] Michael A. Slote, *Morals From Motives* (Oxford: Oxford University Press, 2001).

[40] In the fullest statement of the view, Zagzebski stresses that God is a limiting case of the goodness of an exemplar. Linda Zagzebski, "Exemplarist virtue theory," *Metaphilosophy* 41, no. 1 (2010); *Divine Motivation Theory* (Cambridge: Cambridge Univeristy Press, 2004).

like the Ramsey-Lewis method in order to go from intentions to behaviours, because the intentional and behavioural profiles stand in a process-product relationship. When starting from behaviours we needed an intricate mechanism to derive the process that would produce them, but when we start with the intentional profiles we merely need to refer to their products. This is like the difference between designing a process to, say, pump water, and the undoubtedly easier task of seeing that an Archimedean screw is an effective water pump. This point goes through whether we take the intentional profiles of v-acts as primary, as Slote does, or the intentional profiles of v-traits, as Zagzebski does.

xxvi. Comparing my approach to Jackson's moral functionalism

It is worth comparing and contrasting my approach with the most prominent attempt in the literature to use functional definitions for evaluative terms, that by Frank Jackson.[41] Jackson uses Ramseyfication to provide a schema for how to find natural properties (i.e. ones that fit in a physicalistic understanding of the world) which serve as the extension of evaluative terms. On Jackson's account, evaluative terms are like the natural properties that are the realisors of the moral theory that arises if we take evaluations to be the tacit introduction of such a theory.

I have no quarrel with Jackson's approach up to this point, but his concerns aren't mine. I am not here trying to discover the metaphysics of evaluative properties. All that I need is that whatever the evaluative properties turn out to be, they have the various functional relations I identify. For my purposes, Jackson's sketch of evaluative terms is very much underdescribed, because he goes directly from our everyday moral evaluations to putative moral properties. On Jackson's telling we simply (as much as any of this is simple) take a conjunction of all our moral tautologies and see which properties count as the realisors of the resulting Ramsey

[41] Frank Jackson, *From Metaphysics to Ethics* (Oxford: Oxford University Press, 1998). I take this presentation to supersede the earlier (and less ambitious) version developed along with Philip Pettit in Frank Jackson and Philip Pettit, "Moral Functionalism and Moral Motivation," *The Philosophical Quarterly* 45, no. 178; "Moral Functionalism, Supervenience and Reductionism," *The Philosophical Quarterly* 46, no. 182.

sentence. But Velleman likely will object to such a move because what is under dispute is exactly whether there is something like a consistent conjunction of everyday moral evaluations available. On Velleman's view there would be at best a different theory for each different society referring to that society's schema of action-types and the respective evaluative standards.[42]

To start to respond to Velleman's worry, we must point out that Jackson isn't committed to only doing the most straightforward Ramseyfication from everyday evaluations to moral properties. Going directly from a set of O-terms to its realisors without providing some intermediary theory would make sense if consequentialism were true, and our evaluations are realised by whatever qualities make one set of consequences better than another.[43] But Jackson doesn't want to assume consequentialism for his meta-ethical theory, and rightly so. For one thing, the claim that our everyday evaluations are uniquely realised only by states of affairs and the consequences they represent is patently false, as is admitted by consequentialists as they embark on the project of redescribing and revising our everyday evaluations into consequentialist terms. So, it is agreed that we need intermediary theories that our evaluations cash out into, and this intermediate theory in turn cashes out in the realisor properties (physicalist properties, if the kind of naturalism Jackson defends turns out to be true). Since the v-types are omnipresent in everyday evaluations, they are one fruitful avenue to pursue to see what kind of intermediary theory can be harnessed.

[42] In this respect I expect Velleman's relativism to be much like Gilbert Harman's, which emphasises the extent to which moral evaluations are theory-laden by way of the different frameworks that linger in the background. Harman emphasises the similarities of cross-cultural evaluations to comparing scientific observations across different theoretical backgrounds. By introducing the Ramsey-Lewis method for theoretical terms to address Velleman's relativism of action-types I've as it were dragged Velleman's Kant-inspired framework into Harman's philosophy-of-science one. What links both their theories is that they identify relativism of action-types as underlying moral relativism. See Gilbert Harman, "Moral relativism defended," *Philosophical Review* 84, no. 1 (1975); Margaret Gilbert, "Critical notice: Gilbert Harman and Judith Jarvis Thomson, Moral Relativism and Moral Objectivity," *Noûs* 33, no. 2 (1999).

[43] Perhaps this would only work for direct rather than indirect consequentialism. This makes the next point all the more telling.

XIX. Tying the pieces together

In this concluding section I tie together the conventions at the level of the v-complex with the conventions at the level of individual v-types.

In §III I described a network of v-types as a v-complex that is a matter of convention, and in §IV I decomposed our descriptions of v-types into behavioural and intentional profiles, using these in §V to link action-types to trait-types. It is a matter of convention, because what v-schema to adopt is an SUP case. It is strategic, because we are talking about interpersonal patterns of behaviour, including the recognition of certain kinds of action as being approved of in a particular v-schema; the expectation about how you and others will behave given the adopted v-schema; the expectations about what is accomplished by acting in the expected way and those expectations' knock-on effects on means-end reasoning; and so on. It is underdetermined, because by common assent the evaluative points of the v-terms don't specify a single correct v-complex (to put it in my terms).As I've argued, the adoption of a v-complex is a limited convention in response to this SUP case.

As I've highlighted throughout the chapter, action-trait pairs aren't self-contained. That means that if we have a conventional specification of some specific v-type, be it an action or trait, that will not only settle what the respective action or trait is, but also partially settle what the related v-types are like. In §V.iii I took a given specification of an action-type, courage as it pertains to samurai, what they called *yū*. Earlier I had also discussed courage as it pertains to *condottiere*, what they called *ardimento*. I then stressed that both *yū* and *ardimento* have the same evaluative point: regulating your response to danger when it threatens something of value. I then tracked the way we expand on the v-type *yū* to encompass not only some set of behaviours but also some set of intentional features and character traits. It goes without saying that because the behaviours involved in *ardimento* are different, the respective intentional profiles and v-trait will be different as well, and this is amply evidenced by contemporary accounts of what *ardimento* involves. The condottiere is called upon to be daring and

audacious, cunning and effective both on the battlefield and in nurturing their alliances. These characteristics feature only marginally or not at all in what is called upon for samurai. Nonetheless, for the *condottiere* they are all part and parcel of appropriately responding to danger, and fit into the historical contingencies of their situation, just as the features of *yū* fit into the contingencies of feudal Japan.

The same framework is going to apply to Velleman's examples. *Vranyo* is going to find a place in the Russian v-schema and *étok-étok* in the Javanese one, being part of the conventional approach to issues regarding truth-telling and what Hume calls the virtues of cheerfulness, wit, and charm, such that it is understood that the extravagant but straight-faced bluffing that constitutes *vranyo* and the harmless mischievousness of *étok - étok* carries no expectation of truthfulness and serves to advance good cheer. *Kala:m* and *kizb* will play part of the conventional approach to expectations of truthfulness and respect for the social standing of certain members of the community. And so on.

The above is an instance of nested conventions, as discussed in Chapter 2. The implementation of a v-complex is the wider convention, covering the whole network of evaluative terms that are recognised as virtues and vices. The narrower conventions are those that regulate the individual v-types. For a samurai to respond to danger appropriately—the evaluative point, and the most proximate purpose—also involves the more distal purpose of conforming to *yū*, as well as the more distal still purpose of upholding the v-complex of feudal Japan; for *condottieri* to respond appropriately to danger has the more distal purposes of conforming to *ardimento* and of upholding the v-complex of late-medieval Italy. In this way conventions permeate the evaluative framework of samurai and *condottiere*, and everybody else, and this complicates but doesn't undermine the possibility of understanding or evaluation across societies.

5. Knowing and Unknowing Rightness

I have throughout the thesis presented the view that there are many moral questions of import where we should depend on conventions, especially limited conventions, for guidance. There are objections to this approach on two fronts that this chapter is meant to address. The first kind of objection is that it often doesn't seem like we're guided by conventions when we engage in moral reasoning—as we saw in Chapter 2, Brennan, Eriksson, Goodin, and Southwood make repeated appeals to this in their objections to conventions being normative. The second front is that there is a paradigm of moral action, made explicit in theories like those of the Stoics and of Kant, where a right action is one where (amongst other things) the agent can produce a chain of reasoning in support of that action that draws its justification from sound moral principles. In this chapter I address both these fronts at once, by highlighting how the ability to succeed at attaining some moral end can come apart from knowing where this ability comes from or why it is appropriate. This means that objections like those of Brennan, Eriksson, Goodin, and Southwood fail because (*pace* Lewis) there should be no general expectation that someone who follows a convention know they are doing so; and that being able to produce a chain of reasoning in support of the action you perform is one good way to attain moral ends but not the only way.

In this chapter I introduce a distinction between *mere* and *knowing conforming*. Knowing conforming is when you not only do what is appropriate but can also reproduce the reasoning that vindicates that act as the appropriate one. Mere conforming is when you cannot but instead either reproduce faulty reasoning or can't provide any at all. Knowing and mere conformity form a spectrum: someone like the Stoic conception of a sage, who has a perfect understanding of all the factors that go into making a particular course of action right and acts in the recognition of those factors, would be at the extreme of knowing conformity; whereas the

moral equivalent of Blockhead who acts like a normal person would but because it is specifically instructed to do those things in every detail would be at an extreme of mere conformity.[1] Most of us will fall somewhere in the middle, where we recognise and act in recognition of some of the features of the case, whereas where other features are concerned we don't appreciate their import but manage to do what is right nonetheless.

The point of this distinction is that in some circumstances, like when people act according to a limited convention that supplements their moral principles, your ability to do what is moral can survive even in cases where you are wrong about what it is that makes those actions moral. While in keeping with the rest of the thesis I focus on conventions providing moral guidance, the point is general across any instance of action-guidance where conforming to a regularity is an efficacious way to achieve the end of the activity. That means it also applies to cases where our action-guidance doesn't come from conventions, and also to non-moral cases, like knowing how to bake bread or how to navigate through traffic. The moral case is not the only case, but it is the most contentious one, and it is fair to say that it is also the most difficult and most interesting case.

XX. What we learn from action-guidance

What concerns us here is the epistemology of action-guidance, and the epistemic features of action-guidance we receive from conventions and other social regularities. With 'action-guidance' I mean the what is provided to agents by commands, advice, prompts, suggestions, and other ways of getting to know what you should do. The expectations that are the constituents of social regularities also provide action-guidance, since in regularities these tell you what you should do: in these situations, you should act as expected. To investigate the epistemology of action-guidance, I introduce a distinction that is general to action-guidance,

[1] See Ned Block, "Troubles with functionalism," *Minnesota Studies in the Philosophy of Science* 9.

whether its source is a social regularity or not: the distinction between *first-order* and *higher-order knowledge of action-guidance*. First-order and higher-order knowledge are when you know, respectively, the first-order and the higher-order features of some piece of action-guidance. Let us consider these in turn.

The first-order features are the ones that feature as the objects of that piece of action-guidance, and their relevant properties and constituents.[2] For instance, for the suggestion 'you should close the door', the first-order features are the door that should be closed, that you close it by swinging it on its hinges, that it counts as closed when the latch has engaged with the door frame, and so on. To put it precisely, consider what is sometimes called the 'satisfaction conditions' of a piece of action-guidance, being the proposition that expresses the state of affairs that results when the action-guidance has been followed—the first-order features are those that feature in that proposition at the relevant level of detail.[3] First-order knowledge amounts to the ability to correctly identify these features.[4] Examples of first-order ignorance would be if you don't know which door to close or what counts as closing the door (say, if it has an unfamiliar latch system).

In contrast, the higher-order features are those that aren't themselves objects of the action-guidance, but are instead features of the action-guidance itself. This will include trivial linguistic features, like that 'you should close the door' consists of five words, but more pertinently, it will include things like what the reason is you want the door closed, why you expect the other party to do as you ask, the reasoning you went through which resulted in you

[2] Here we have the same problem as we do whenever we try to determine what is expressed in a sentence. It seems overly restrictive only to count those things explicitly mentioned, but it's hard to draw the line at what else counts as expressed. Some things will count as expressed because they are entailed by the things that are explicitly mentioned. Does everything entailed by the things explicitly mentioned themselves count as expressed? Probably not. I mention this problem in order to set it aside.

[3] The satisfaction conditions for action-guidance may be path-dependent: for instance, you can command your child to not make sure that the rubbish has been put out, but that they should do it themselves.

[4] For my purposes it doesn't matter whether this knowledge is knowledge-that or instead some kind of non-propositional knowledge-how, and for simplicity's sake I will treat them as instances of knowledge-that.

issuing the action-guidance, and so on. For instance, that closing the door will lead to the room being less draughty would be a higher-order feature. It isn't a first-order feature because someone can follow the action-guidance successfully and it nonetheless not make the room less draughty, like if the door isn't effective at stopping the draft or you mistook what causes the draft. Higher-order knowledge is what results if you have cognitive access to these higher-order features, say by successfully deducing that you asked me to close the door in order to stop the draft. An example of higher-order ignorance would be not knowing this, as would be indicated by the comical situation where I close the door and then immediately go on to open a window in order for there to be a breeze.

My claim is that much of the time when we conform to action-guidance, we may succeed in doing so despite suffering from *higher-order ignorance*: the failure to recognise the moral import of some salient feature of the situation. We will start with an example from a simpler, non-moral domain, and then make use of the same points for addressing moral guidance.

Consider the knowledge required to be a baker. The relevant action-guidance is the training the baker was received for baking bread. It is one thing to know a workable recipe for baking bread, and it is another to understand the various chemical processes involved which turns a lump of dough into bread. We don't in general expect bakers to also be scientists: we only want them to be able to put into place the procedures required to bake the bread; we don't also ask them to be able to explain the intricate chemical reactions that they are harnessing. Here the first-order knowledge the baker displays are brought to bear on the various procedures involved in baking bread—mixing an appropriate dough, kneading it, letting it rise, baking it at the right temperature for the right amount of time, and so on. All of these are the first-order features: the things that appear in the satisfaction-conditions of the action-guidance. All of the chemistry that makes this an effective process for baking bread counts as higher-order features, as does the history of this process, who taught it to the baker, and so on.

Let's concentrate on the kneading. The baker knows an effective method of kneading the bread, like stretching the dough and pounding it out by hand, which they then put into practice. This exhausts what the baker needs to know in order to succeed in kneading the dough. Notice that they do not need to know why it is that kneading has these effects. They needn't know about how the repeated manipulation of the dough leads to the forming of gluten chains that trap the carbon dioxide bubbles released as the dough warms during baking. The first-order knowledge is knowing to knead the bread by doing such-and-such; the higher-order knowledge is knowing why kneading accomplishes what it does.

I'm not saying that the baker is entirely ignorant of the higher-order facts—knowing that failing to knead the dough properly will lead to the bread collapsing when baked is a piece of higher-order knowledge.[5] Instead, I say that the baker is ignorant of much of the important higher-order features of the action-guidance they are following, certainly to the extent that they couldn't give a compelling description of what it is that they are doing other than reporting that this method succeeds in producing good bread.

In baking and many other domains including, as I argue, some instances of moral guidance, we can have the first-order guidance and yet have higher-order ignorance. The inability to provide a chain of reasoning that starts with sound principles and concludes with the action-guidance in question doesn't harm the standing of the action-guidance the baker makes use of. Accordingly, if what we care about is whether people succeed in doing what is moral, there may be circumstances where they can reliably and repeatedly accomplish this without any explicit knowledge or understanding of the sound principles that vindicate their action, just like bakers don't automatically also understand the chemistry.

As we discuss our baker's understanding of what they are doing when baking bread, we contrast their higher-order ignorance about why it is that kneading the dough, etc., is

[5] In Chapter 6 we look at what may happen if someone suffers from total higher-order ignorance.

efficacious with their knowledge that the process has the desired effect. The knowledge they do have is enough to achieve what I call *first-order success*: to meet the satisfaction-conditions of the action-guidance. Individuals attain first-order success through performing such-and-such actions; the action-guidance gives the criteria for achieving that success. The purpose of the baker's training isn't just to gain some piece of knowledge, but to have that knowledge accomplish something—in our example, baking bread. This isn't an epistemic category, as we can see by comparison with cases where people have first-order knowledge but can't reach first-order success. If our baker was cut off from a supply of yeast, then all their expertise regarding getting bread to rise won't do them any good, because they aren't equipped to make use of it.

The interesting phenomenon that this chapter is devoted to is how we can have first-order success accompanied by higher-order ignorance, even in normative and morally loaded domains. We may be unperturbed by a baker's ignorance about why they bake bread the way they do, but according to a paradigm of moral reasoning we should expect someone to be able to offer a chain of reasoning starting in sound moral principles

xxvii. Higher-order ignorance and paired profiles

The notion of first-order success I've spelt out above ties in with the paired profiles approach developed in Chapter 4, where we describe actions by way of a pair of descriptive profiles: a *behavioural profile*, containing the bare descripting of the behaviours involved in the action; and an *intentional profile*, containing the pertinent psychological features of the action such as its motive and the kinds of perceptual sensitivities involves in it. Something counting as a first-order success means that it displays a particular behavioural profile, the one that counts as an instance of the target action. Failures to attain first-order success are when the behavioural profile of the action fails to count as an instance of the target action. For instance, if the target action is to bake a loaf of bread, then the first-order success is to transform a collection of

ingredients (flour, yeast, salt, water, etc.) into a loaf of bread. The behavioural profile is to mix the dough, knead it, let it rise, warm the oven, oil the pan, bake the bread, etc. Another example, this time of a social situation, is navigating a traffic intersection. The first-order success is to make it through the intersection in the normal way without causing a disruption: stopping at the stop sign, conforming to the right of way rules, promptly taking your turn to cross, and so on. This just is the behavioural profile of making it through the crossing in the normal way. To talk about the first-order success is to highlight the end of the target action; to speak of the behavioural profile is to highlight the means; but they are just different descriptions of the same terrain. To bake some bread just is to follow the recipe; to appropriately cross an intersection is just to follow the appropriate road rules.

What then goes into the intentional profile? When we bake bread or drive a car there are things we keep in mind—respectively, a recipe and the road rules—but it's important to stress that keeping these things in mind is at most a small part of the intentional profile. This is where the distinction between mere and knowing conformity comes into play. To knowingly conform to something is to have a full grasp both of what is required to succeed at it (knowledge of the first-order features) and also of the pertinent background features such as what principles require this action to be performed (knowledge of the higher-order features), and to conform under your own self-control. This notion has obvious relevance to Stoic views on ethics and self-control, where the paradigm of a good moral agent is the sage, someone who only assents to beliefs and only performs actions when they are secure in knowing that it is right to do so. It is also what I take Kant to be after when he argues that autonomy and conforming to the moral law go hand-in-hand. These are among the most explicit references to the paradigm of moral action I'm using as a stalking-horse in this chapter. In contrast, to merely conform is to do as the action-guidance recommends without an accurate grasp of the higher-order features of the action-guidance. The point of this distinction is that first-order success requires only mere

conforming. So, to be a knowing conformer is to not only have the right behavioural profile but also the right intentional profile.[6]

The question then is: what is the intentional profile of mere conforming? People who merely conform to the specified action are of course able to reliably conform, just as bakers managed to bake bread with yeast for thousands of years before the chemical processes in question were described. Reliably conforming to some piece of action-guidance (the bread recipe, the road-rules of your community) requires no more than reliably recognising that the action-guidance is in effect and reliably performing the specified response.[7] But here is the crucial point: these first-order requirements—recognising the action-guidance, knowing the required response, and actually responding—are entirely neutral regarding the source of the guidance. However it may be that the individual got to have this first-order guidance and became disposed to respond in the specified way, it will suffice for first-order success. This counts even if people have patently false higher-order beliefs about the first-order guidance.

XXI. First-order success despite higher-order ignorance

We now have all the tools we need to analyse cases where someone attains first-order success accompanied by higher-order ignorance. To make the point especially clear, I will give an example where the agent is not only ignorant of the relevant higher-order features of their actions, but are actually in error about them. Consider this fanciful modification of the Good Samaritan scenario. A Silly Samaritan comes upon an injured man on the side of the road, and takes great care to help him. The Silly Samaritan helps the injured man to safety, sees to his accommodation and care, and is genuine in his concern for the injured man's well-being.

[6] This means that it's the more general version of the acting according to virtue vs. acting from virtue distinction discussed in Chapter 4.

[7] It has been argued that you can do the same with far less. See for instance Ned Block on the behaviour of Blockhead, Ruth Millikan on the regularities of hoverflies and Brian Skyrms on the communication of bacteria. Block, "Troubles with functionalism."; Millikan, "Truth, rules, hoverflies."; Skyrms, *Signals*.

However, the Silly Samaritan does this because he thinks the injured man is a leprechaun, and that he will reward the Samaritan with a pot of gold. The Samaritan may be happy to see the injured man on the mend in its own right, but delighted with the prospect of the reward. The beliefs of the Silly Samaritan are daft—to add insult to injury, he is even mistaken about how leprechauns behave, not even managing to be right about the storybook element of his plan. And correspondingly the Silly Samaritan will be sorely disappointed in time. But none of this threatens the first-order success of having seen to the care of the injured man.

xxviii. The alternative method model

The Silly Samaritan may by chance be a boon to his neighbour, and similarly the efforts of pantomime villains may be so self-defeating that they end up being for the good, but there couldn't be any expectation that such acts reliably have good ends. These examples are of one-off first-order successes, but we are also concerned with repeated successes, especially where conventions are concerned. To give a general description of how first-order success despite higher-order ignorance is possible, I offer what I call the *alternative method model* (AMM for short): in some cases there are multiple procedures that would suffice for attaining first-order success, and in those cases knowledge of one method allows ignorance about a different one. The point of this is that someone can attain first-order success through one method, but lack any grasp on why that end is the one they should be aiming for, because that justification is tracked by a different method to the same end.

Consider how a child may have two different ways by which to know to avoid being burnt by the stove: by the one method, the child may appreciate how hot the stove can get and that the stove will damage their skin and flesh when heated; by the other, the child may know that their parent has told them to avoid the stove. First-order success in this case is not getting burnt. The behavioural profile of this success is avoiding the stove when it is hot, being careful not to

touch it when nearby, and so on.[8] But their intentional profiles will be very different, since the child who understands the dangers of the stove will respond directly to those dangers, whereas the child who responds only to the parents' warning will be responding to features of the parents, not the dangerous features of the stove.

Here are the pieces of the AMM. Firstly, there is some given normative framework that provides us with action-guidance. In the case of moral reasoning this is likely to be some set of principles, but something as workaday as 'burns are one kind of harm to avoid' is a perfectly good example. Secondly, there is the end that is recommended by that framework within some given situation. Call this the *desired end*, since it is the most natural way to refer to the end that someone aims at is by way of their subscribing to the framework in question. In our example, the desired end would be not getting burnt by the hot stove. Thirdly, there is the action-guidance that results from that framework in some given situation. In our example, this is 'avoid a burn by not touching the hot stove'. Call this the *privileged method*. It is privileged because it is the method that captures why it is that the framework recommends the desired end. It is what you know if you are a knowing conformer. But there may very well be an *alternative method*, such that conforming to that method also reaches the desired end, but the action-guidance is genuinely different, in particular by referring to different features of the situation than those picked out by the privileged method. In our example, the alternative method would be for the child to obey the parent—since the parent has ordered the child not to touch the stove, this also reaches the end of the child not touching the stove. Notice that this action-guidance makes no reference to heat or burns. Nonetheless, given what the world is like, avoiding touching the stove is also to avoid burns—the desired end.[9] These methods, privileged

[8] What the child reports when asked about the reasons not to touch the stove will be a behaviour in its own right, but not one that contributes to the first-order success, and thus not part of the behavioural profile of that success.

[9] This links up with the discussion of nested purposes in Chapter 2. Just as the child's obeying the parent is an efficacious way to avoid burns, so too for Millikan's examples—a circus dog's riding a bike is an efficacious way for it to get food and survive, a hoverfly intercepting moving dots in its field of vision is an effective way for it

and alternative both, may produce both one-off actions (like in the Silly Samaritan example) or repeatable ones (as in the child-and-stove example).

The alternative method model links to higher-order ignorance in the moral domain in that we can learn to do something by way of a method that is perfectly capable of reaching first-order success, but which isn't the method that links in the appropriate way to sound principles of moral judgement. In such cases we reach the end that is recommended to us by some piece of sound moral reasoning, but which recreates only the end and not the reasoning. In our example above the child reaches the first-order success of not getting burnt by following the parents' advice, even when entirely ignorant of the bad effects that the hot stove would have on their skin and flesh. But of course it is unsurprising that someone may not know every feature of every possible method for attaining success at their target act. Some of these methods may be especially pertinent in some or the other context, such as being easy to teach and remember, such that we may end up using them as the currency of our action-guidance, even if they aren't the privileged methods. This would lead to cases in the moral domain where we have first-order success accompanied by higher-order ignorance.

The above kind of case may tempt one to say that the point of warnings and advice is to provide the audience with an alternative method of attaining first-order success at the target action. This may be because the advice is meant to stand in for the reasoning of the audience (as in the case where the child is warned away from the hot stove), or as a reinforcement of their own reasoning (as in the case where you encourage someone to do something they have already at least considered doing). You may say the same even of orders and commands, at

to mate with female hoverflies and reproduce the species. The same goes for Marmor-style deep conventions: conforming to the surface convention of playing chess is an effective way to conform to the deep convention of playing a game of strategy. Here we don't just note that there are multiple ends available for a given action, but we highlight one of them as privileged, and the action-guidance which refers to that end as the privileged method. In other respects the story here is just the same, up to and including the point that the pursuit of one purpose out of the telescoping series is intentional only under a description. See Millikan, "Truth, rules, hoverflies."; Marmor, *Social Conventions*.

least under something like Joseph Raz's service conception of authority. I will develop this point no further here than by noting that in Chapter 3 I provided an analysis on which many commands do provide such an alternative method, by way of creating conventions to do such-and-such in the specified circumstances. That analysis was meant to show that in the surveyed cases conventions would be a possible and efficacious method for providing such alternative methods for what I have here called first-order success.

To illustrate how the AMM can be used to understand views developed in moral philosophy, here is how we can understand indirect consequentialism as an application of the AMM. The consequentialist believes that all moral evaluation is evaluation of the consequences of action; the indirect consequentialist believes that the best global strategy for attaining good consequences is not to aim directly at these consequences on a case-by-case basis but to put in place general measures that have the best overall consequences, even if they don't have the best consequences in every particular case. On this view, while the privileged method would be to pursue a course of action because it has the best consequences, the alternative method would be to pursue some strategy that isn't directly consequentialist. So, any instance of an indirect consequentialist strategy that succeeds in maximising overall utility in the long run would be an instance of the AMM in action. Correspondingly, the difference between mere vs knowing conformity would be that the mere conformer doesn't know how the particular instance of the AMM brings about the best consequences overall, whereas the knowing conformer does.

xxix. Higher-order ignorance and Lewisian conventions

On the Lewisian account that I'm building on, the first-order content of a convention is the expectation that everybody does φ, everybody expects everybody else to do φ, and so on. Crucially, this expectation is neutral as to how it is formed. It is the agnosticism about the source of the expectations that allows conventions to perform the task Lewis wants them for: to allow the kind of cooperation you often get by way of agreements without requiring people to

make explicit agreements. But this agnosticism has other effects as well. Notice that the details of what cooperating in each instance consists in—what the regularity is that people conform to—counts as first-order action-guidance, whereas the source of the expectation is among the higher-order properties of the regularity.

While the requirements for knowing that some specified action-guidance is the object of a convention are very strict (e.g. the demanding standard of 'common knowledge'), there are almost no requirements regarding the higher-order properties of conventions: only that each agent must correctly ascertain that everybody else has the appropriate expectations, and from those expectations the conforming behaviour is meant to follow. This is a relatively modest requirement. It is certainly far less demanding than the common view that to be able to be secure in your judgement to do such-and-such you must have access to a justification making reference to sound principles. But Lewis's achievement is to show that it is enough to secure reliable cooperation of the sort that is ubiquitous throughout our social lives. Conventions are thus a striking example of first-order success that can accompany higher-order ignorance.

I take this point further than Lewis does. Lewis has as one of his criteria for a convention that the participants know that there is some other possible regularity they could have coordinated towards instead. Knowledge about other possible regularities counts as an instance of higher-order knowledge, and is on my analysis something that the regularity can survive without. This kind of higher-order knowledge is one avenue to conforming to the regularity—something knowing conformers would have—but it is by now commonplace that conventions can survive without, as argued by Tyler Burge.[10] On this point I side with the critics against Lewis, but this doesn't diminish the import of Lewisian conventions (and limited conventions in turn), because it shows that they are more robust than Lewis gave them credit for. In effect Lewis didn't take his observation that conventions are neutral with regards to their sources far

[10] Burge, "On knowledge and convention."

enough.[11]

XXII. Why mere conformity is second-best

In this section I lay out all the problems with depending on mere conformity for guidance, and then in §IV I show that even if we take all these problems into account mere conformity and the AMM are nonetheless provide worthwhile action-guidance. So, here I discuss all the reasons you may have to worry about mere conformity, and later I show why these worries don't undermine its moral import.

It is easy to imagine how higher-order ignorance can arise in a community that has regularities as a part of its established moral framework. Here is one likely scenario. A population of conscientious individuals have a grasp on their moral principles but run into SUP cases. In response limited conventions arise—it doesn't matter whether they do through happenstance or coordinated effort. The regularities selected by the limited conventions then become an established part of the moral framework of that community, and they depend on it for guidance. As people start conforming to the regularity, it makes way for explicit recognition of the principles behind the limited convention to diminish. This allows higher-order ignorance to arise. Below I look at the problems that may arise with it.

Conforming to a regularity allows you to reliably and spontaneously reach a benign outcome in the situation that the regularity addresses. But this doesn't exhaust what is required by ethics—crucially, mere conformity to a regularity gives no guidance for handling novel situations, nor for comparing different possible courses of action against each other. It only gives superficial guidance for evaluating the regularity you conform to, based on your acquaintance with the outcome it secures (acquaintance which is likely to be incomplete in any case)—as discussed above, knowledge of higher-order features isn't secured by this

[11] I develop this theme in Chapters 2 and 6.

acquaintance. Going further with investigating the shortcomings of mere conformity, a difference which till now has been of little importance starts to become operational: the difference between ignorance about higher-order properties and having false beliefs about them.

It is useful to stress again that there are many different kinds of higher-order features of a piece of action guidance you may be ignorant of. The clearest example of this is to not know whether a given course of action is consistent with the principles that underlie your moral framework. This is the feature I use for most of my examples that follow, but it isn't the only one available. Other pertinent higher-order features are: knowing what first-order features are the ones that make that action lead to a benign outcome; whether there are alternative actions that are also consistent with the principles; or knowing whether the regularity is uniquely determined by the principles or conventionally selected; and so on. The fact that the most obvious higher-order feature of note is whether the course of action is consistent with the principles shouldn't hide the fact that there are other features available, and sometimes they make a difference (as they do in one of the examples below).

xxx. The shortcomings of mere conformity

In the face of higher-order ignorance an individual who considers some situation that isn't the subject of a regularity is in a worsened version of the position someone is in when facing the *strategic underdetermination problem* (SUP for short), which was the topic of Chapter 1. In SUP cases, there is a range of options available, but not the resources to pick one out as the appropriate thing to do, since the principles used to determine what is appropriate underdetermine what option to take. In these cases you also need to be able to predict what the people around them will do, since the outcome of your action depends on what they will do, but since the principles are underdetermining you don't know which of the available options they will take. Accordingly, your uncertainty about their actions bleeds over into uncertainty about

what you should do.

Insofar as the individual has higher-order ignorance about the content of a principle, their reasoning will have even less traction than someone when their principles underdetermine what they should do. For instance, not only would the extent of indeterminacy faced by someone subject to higher-order ignorance be greater because they would have to consider options that are (unbeknownst to them) inconsistent with the principles, but they may be uncertain about what to do even when (unbeknownst to them) the principles determine some uniquely appropriate option. They may even come to unknowingly try to bring about what I called malignant outcomes in Chapter 1, outcomes that are properly worse by the lights of the principles than some other live option. So, the individual suffering from higher-order ignorance has the harms of uncertainty, and the extra danger of unknowingly adopting a course of action that leads to a malignant outcome.

The same problem is also in effect in SUP cases, further exacerbating their difficulties. In those cases, not only is the range of underdetermination faced by a given individual in question (itself widening the range of options the individuals don't have the resources to decide among), but the various different kinds of possible ignorance distributed across the community plays a part as well. I'll illustrate this with a toy example. If there is a situation with four possible responses {A, B, C, D} and the principles in play actually exclude {C, D}, then we are left with a SUP case where the individuals can't choose between A and B, which is bad enough. But if one individual X in the case is ignorant of the principles excluding C, then X is stuck with an underdetermination between {A, B, C}, and the other individuals in question are faced with the prospect of X pursuing C. Thus the SUP case now involves a choice among A, B, and C, even though C is a malignant outcome. The fact that the other individuals may not know about X's ignorance makes it worse: now there is the further question of whether they believe they are in the {A, B} version of the situation, or the {A, B, C} version. This is a further barrier to

cooperation.

Similarly, if there is a different individual Y who knows that C is excluded, but is ignorant about D being excluded, then Y's reasoning is underdetermined between {A, B, D}, which means that the SUP case is now between all four possible outcomes {A, B, C, D}, despite the latter two being excluded by the principles. Thus, all the harms of SUP cases that arise from ignorance about what the other party will do also affect the person subject to higher-order ignorance when they face situations not captured in a regularity, and to a greater extent.

The position is distinctly worse for someone who is not only ignorant of the higher-order features, but positively mistaken about them. There are two dimensions to this. Firstly, they face a worse version of the strategic underdetermination problem that was discussed above for cases of higher-order ignorance. The reason why it is worse is because now not only are they in danger of not excluding options that are excluded by the principles, but are in danger of excluding options that are allowed by the principles and should be considered. It is perfectly possible that in this way someone suffering from higher-order error may entirely rule out all the options that are consistent with the principles and leave themselves only with a choice between options leading to malignant outcomes.

Let us construct another toy example to make this point. Again we have a situation with four possible responses, {A, B, C, D}, and an individual Z who subscribes to a moral framework where the principles exclude {C, D} from consideration, so there's an underdetermination problem. Suppose that Z suffers not just from ignorance but positive error about the pertinent higher-order features, and that Z's false higher-order beliefs imply that Z believes that the principles exclude {A, D}, so Z understands the choice to be between {B, C}. This is worse than the position X or Y was in when suffering from just ignorance, because Z has excluded from consideration an option that would be allowed by the principles. Since it has to be at least partly by accident that X or Y arrives at a benign outcome (the extent of the error

increasing how accidental the choice of a benign outcome would be), the likelihood of such a happy accident decreases the smaller the proportion of options leading to benign outcomes are compared to those leading to malignant outcomes.[12]

xxxi. Higher-order ignorance and faulty extrapolations

In addition to problems involving decisions within a particular recurring situation, there are also problems that arise when you range across situations. Here again being subject to some positive error is a distinctly worse position to be in than just being ignorant, because someone who is just ignorant thereby loses one kind of grip they may have on the case, whereas someone who suffers from a positive error will find themselves wading ever deeper into the weeds.

If you just lack a certain piece of pertinent higher-order knowledge, then you can't use that knowledge to inform your other decisions. For instance, mere conformers can't tell which of the features of the given course of action lead to it being selected by the regularity. For each feature, that feature may be one that is relevant to the principles (say, if it's a principle not to lie, the fact that the course of action is to tell the truth), or it may not play any role in the determination of that course of action as appropriate. Without a full knowledge of the principles, you won't be able to do more than make a half-informed guess at best. A mere conformer doesn't have full knowledge, and as the extent of their higher-order ignorance grows, their ability to make good guesses on which features are pertinent diminishes. This matters for evaluation across different situations, since they can't conclude that consideration of some particular feature of that case carry over to a different recurrent situation. That is, higher-

[12] If we add another individual who is subject to a different error such that that individual believes the choice includes the options {B, C} (and perhaps others as well), then we'll again have an SUP case across all four options {A, B, C, D}. This just indicates that in both the higher-order ignorance and error cases we can easily construct situations where the individuals involved have lost any traction for deciding between the available options.

order ignorance strips from you the ability to extrapolate from settled cases to unsettled cases.

Someone who suffers from higher-order error, and not just ignorance, suffers from a worse version of this inability to extrapolate correctly, because insofar as they assent to false beliefs about higher-order features of some instance of action-guidance, they will assent to extrapolations that are the consequences of these false beliefs. Insofar as these extrapolations lead away from benign outcomes towards malignant ones, someone will be harmed by assenting to them. Call this the class of *faulty extrapolations*. Someone who is just ignorant avoids extrapolations, faulty and correct alike; someone who suffers from higher-order error will by that token be vulnerable to accepting faulty extrapolations.

xxxii. An example of faulty extrapolation and its harms

To give an example, consider the widespread and generally appropriate instruction to respect your elders. Let us suppose, as is plausible, that this instruction is appropriate because following the guidance of your elders is a good strategy for conforming to established practices, and conforming to established practices is in turn a good strategy for achieving good and avoiding ill. In this case, the pertinent higher-order features of the instruction to respect your elders include that elders are good sources of knowledge of established practices (having participated in them for a long time), and that the established practices are established because they are generally favourable. However, it is commonplace for young people to suppose that the reason they get instructed to respect their elders is because whatever an elder instructs you to do is right thing to do. This is a faulty extrapolation, for familiar reasons.[13]

[13] This is of course an example of the kind of relationship to authority that is the subject of Plato's *Euthyphro*, leading to a dilemma where either the authority would then be unconstrained in their pronouncements in a way that is offensive to our sense of right, or is responding to some further feature of the case which isn't just their say-so which isn't captured in this formula. While some very few philosophers will bite the bullet and claim that God's say-so is unconstrained, presumably nobody takes this tack for the guidance of their elders. In this case, I specified (as seems obvious in any case) that there is some further feature not captured in the formula—the elder's greater familiarity with the established practices that for the most part help the members of the community. This makes the dependence on the elders' guidance redundant. In my terms, when the elders' guidance leads to a benign outcome, it is an example of the alternative method model in action.

At least two distinct types of erroneous positions can be motivated by this faulty extrapolation. Firstly, this can lead to the naïve but not generally harmful practice of a youngster following the instructions of an elder without question. They will arrive at benign outcomes so far as the elders do in fact instruct them in how to conform to established practices, and these practices are regularities. Someone with this naïve view will be confused about morality and the consequences of their own moral framework, and this is likely to lead to genuine harms in cases where there aren't regularities in effect, or these regularities are being revised or need to be revised, since then conforming to the established practices the elders are likely to point to aren't enough to reach a benign outcome.

Secondly, someone who assents to the higher-order belief 'what the elders instruct you to do is the right thing to do' can also adopt the cynical view that 'right' here means something like 'what is to the elders' advantage', and resent the established practices by that token. Someone adopting the cynical reading then becomes alienated from the benign outcomes the established practices are meant to lead them towards. Then their cynicism leads to the faulty extrapolation that there isn't a reason to conform to these established practices when you don't wish to pursue the elders' interests, meaning they reject the established practices. Insofar as the established practices are effective ways to reach benign outcomes, they harm themselves, and insofar as these practices give guidance in SUP cases, they harm their fellows as well.

XXIII. Why mere conformity doesn't undermine morality

The work in §III may suggest that we should never make do with mere conformity, since it has been declared to be second best. I will now argue that this is not the lesson we should take from the above. Mere conformity is second-best, but it is still good enough. What is more, I will argue this for the case of conforming to moral guidance, which seems to be the most difficult and interesting case. Most of the examples of first-order success accompanying higher-order ignorance I've surveyed thus far (bakers and gluten, navigating through traffic, a child avoiding

a hot stove) have involved action-guidance in situations that aren't obviously morally loaded. This may lead the reader to worry that the AMM may miss something that is special about moral reasoning, and accordingly not be fit for purpose. And there is a principled reason to worry about this: as mentioned, the paradigm model of moral reasoning that has passed to us through the long tradition of ethics is of individuals conscientiously working from moral principles. In this section I address this worry head-on.

There are facets to my defence of the suitability of the AMM to moral cases. The first is that mere conformity doesn't create higher-order ignorance, but instead allows regularities to persist in the face of it. We mustn't confuse something forestalling the harms that arise with some threat as inviting this threat—this would be like blaming glasses for people's vision deteriorating. You may think that the threat would be less likely to arise if there weren't ameliorating factors, but in this case (as in so many others) this turns out merely to magnify harms. The second facet of my defence emphasises that we have a lot of use for mere conformity, especially in social situations where you need many individuals to succeed at the same thing simultaneously. This is especially important with regards to moral education.

xxxiii. Mere conformity only brings benefits

The question is whether a practice that can easily lead to individuals not having a grasp on salient features of their moral framework is by that token inappropriate. Here I argue that it is not, because when the existence of mere conformity makes a difference to how people are able to fulfil the manifest demands of their moral framework, everybody is strictly better off with it. For a community which has mere conformity in it, there is no alternative scenario where the alternative method model isn't in effect and the community is better off.

Consider a population of individuals, each with some given extent of knowledge about the principles and other higher order-features of their moral framework. In the absence of the AMM, the only moral ends that the individuals can achieve are ones that they are able to reason

towards on their own power (that is, that they can display both the right behavioural and intentional profiles). Furthermore, the AMM isn't in effect if the distribution of higher-order knowledge is such that every individual is a knowing conformer, or it may be that there are no regularities or other avenues for the AMM in place. This is because in these situations there is no example of first-order success despite higher-order ignorance. Accordingly, these possibilities say nothing about the effects of the AMM. What's left are the situations where mere conforming makes a difference in the options available to people. This difference can be in one of two ways: people conform to a proper regularity which has a moral end as its first-order success, or conform to some malformed regularity which doesn't have such a moral end. If mere conforming leads them to some moral failure, then that is a shame but they weren't equipped to do any better without mere conformity. Their behaviour may be different, but the result is the same in kind: moral failure. But when mere conformity has as its first-order success some moral end, we reach the crucial point: in those positions where the AMM makes a difference in the kind of outcome reached by the individuals, it does so by empowering them to achieve a moral end. The scenario without the AMM has the individuals achieve the ends that their higher-order knowledge equips them to; in the scenario with the AMM they achieve those ends and then the further ones that the AMM allows first-order success despite higher-order ignorance. This means that in a situation where the AMM is available, to insist on only your own higher-order knowledge is to choose failure over success.

Someone may object that describing the outcomes of mere conformity to some malformed regularity and the outcome of not acting from ignorance as being of the same kind is likely to conceal as much as it reveals. It is easy to imagine situations where conformity to some malformed regularity leads to a worse outcome than would result from a lack of coordination; faulty extrapolations like I discussed in §III would be one kind of example. But in response, it's equally easy to imagine the reverse. The possible outcomes of acting without a way of securing

moral ends are largely unconstrained, and there just isn't much to be said about them. Also, we should resist the temptation to think that not acting doesn't have an outcome: for people to not conform to anything in particular is a course of action just like any other (a different course for each combination of actions by individuals) and has outcomes of its own. When you lack a reliable method of attaining a moral end, you're the ethical equivalent of adrift without anchor: there's no telling where you'll end up, and it's unlikely to go well. Accordingly, while I can't claim that any failure to reach a moral end is much of a muchness, I can claim that there isn't enough to go on to try and appeal to systematic differences between various ways of being adrift. As far as individuals can have them in mind as ends, their features are blank and inhospitable.[14]

Someone may object that the issue isn't so much how the AMM influences actions when it secures first-order success, but the general effect of a community depending on moral means they don't understand towards ends they're not sure about. This is an intelligible concern, but it doesn't make a difference. The reason it doesn't is because the alternative to allowing for higher-order ignorance is the same kind of situation as surveyed above as an instance of choosing failure over success. The alternative to accepting mere conformity is not to allow it, which means that individuals are then restricted to only being able to attain moral ends that they can secure through their own higher-order ignorance. And now we get to the same situation as above: refusals to make use of the AMM stops the members of a community from attaining moral ends it is equipped to secure (if their refusal has any effect).

For the purposes of the thesis I'm mainly interested in cases where the AMM is in effect by way of established regularities, but the above demonstration goes further than that. For

[14] There is the point to be made here to the effect of 'better the devil that you know': that there are reasons to prefer a bad situation where you know what to expect to a different bad situation which isn't stable in the same way. Regularities and the epistemic position they place their participants in are of obvious relevance here. I think there is a lot to say about this point, but I don't want to depend on it here. Partly this is because I think the approach taken in the main body of the text is a better one; partly because appealing to the devil that you know has difficulties of its own that I don't want to be bogged down in.

instance, it also works when the AMM is in effect by way of case-by-case guidance, such as a parent prompting a child what to do. In that case, what the child achieves if there is no AMM in effect is a proper subset of what is achieved if the by prompts of the parent puts the AMM in effect, as demonstrated above. There is another point to be made here: instead of seeing the parent and the child as different individuals interacting with each other, we can instead look at groups of people such that the group follows the lead of some individual with the relevant higher-order knowledge (maybe the same individual every time, maybe not). In this case we are aggregating the higher-order knowledge of the group together, so that in the simplest case if any individual has the relevant knowledge then everybody in the groups benefits from it. This is analogous to the demographer's move from surveying individuals to surveying households or some other collection in cases where the aggregated features are more pertinent than those of individuals. There are many social phenomena which look like they are more interestingly modelled in terms of such groups rather than individuals: the actions of families, workplaces, committees, etc. The same analysis goes for these group behaviours as for individual ones: if the AMM is in effect and gets people to change their behaviour, then they benefit from it.

ii. Allowing for mere conformity makes moral guidance more robust

In the above sub-section we looked at the negative case for accepting a place for mere conformity: trying to exclude it makes matters worse without any benefit. Now we go to the positive case: it is an important avenue through which appropriate behaviour becomes entrenched in a community and contributes to the moral education of its members.

Doing the right thing from mere conformity is clearly a second-best way of doing so, but it nonetheless is an effective way to secure moral ends. [15] There are two distinct reasons why this makes a difference for the viability of regularities, one pertaining to individuals, the second

[15] This section ties in to the discussion of acting from vs acting according to virtue in Chapter 4, Section II.2.

regarding the social dimension of action within a community.

The first reason is that there being fewer requirements for an individual playing their part in a regularity has the consequence that the regularity can persist even among individuals anywhere on the continuum of the extent to which they are knowing conformers. This continuum ranges from someone suffering from total higher-order ignorance at one end to someone who is a moral sage and has perfect knowledge of what is right at the other.[16] This means that an individual can make their way along this continuum, from unknowing to knowing, without it endangering their ability to fulfil the manifest demands of morality. This is important for a lot of reasons, but none more so than moral education. Consider the position of children being socialised into a community. There is a lot of ground to cover between being able to recognise a recurring situation and act in some specified manner, and understanding the principles of the moral order you find yourself in. The latter is the paradigm of moral action, and clearly out of reach of children as a rule. But even very young children can learn to do the former, which is all that is required for mere conforming. So, for all the steps of the moral education they undergo (and adults continue) from this very basic step onwards, children can take part in the regularity. While this point is of the most obvious interest for children, there is no principled reason to stop there. Insofar as nobody or almost nobody is a moral sage, everybody or most people will find themselves somewhere short of having perfect higher-order knowledge, but nonetheless succeed in doing their part in the regularities that make up the moral order of their society. There is also the point that acquaintance with actions encoded within and the benign outcome resulting from a regularity is a good means to learn about the higher-order features of that regularity.

The second, social dimension is that if the bar for playing your part in a regularity is too high, then the survival of the regularity will be in danger because too many people will fail to

[16] More on an individual subject to total higher-order ignorance in Chapter 6.

conform. This is especially important in cases where the regularity being conformed to is a response to a SUP case, when it is a limited convention. Then failing to conform to the regularity doesn't only result in the individual in question not reaching a benign outcome, but it also denies the other people in the community the chance to reach the benign outcome because the necessary cooperation fails to eventuate. In that case everybody is dumped into the harms of uncertainty. It is that harm that limited conventions are meant to address, and against which the desire to only engage in knowing conformity offers no defence against.

XXIV. Conclusion

This chapter has dealt with the extent to which individuals can play their part in both moral and non-moral action without fully understanding what they're doing. The basic thought is that when we split the behavioural profile from the intentional one we see that the maintenance of a regularity requires only that individuals display the right behavioural profile. This means that they can do their part and allow the regularity both to reach the end that justifies it and for it to be promulgated if they are ignorant about the deeper reasons for why that is the appropriate behaviour, or even in positive error about those reasons. Not knowing pertinent features of the action that aren't contained in the behavioural profile is what I've called 'higher-order ignorance'. Someone who displays both the target behavioural and intentional profiles is a 'knowing conformer'; someone who displays only the target behavioural profile is a 'mere conformer'. Here I presented the mechanism that allows for this split, which I called the 'alternative method model': there are multiple intentional profiles available that will suffice to produce the target behavioural profile, so we can replace the knowing reasoning of a fully-informed individual with one of these surrogates, including very bare intentional profiles such as 'doing what someone told me to do' to full-blown cases of mistaken confabulation. I surveyed various problems that can arise if individuals suffer from higher-order ignorance, of which the most prominent is that they are liable to faulty extrapolations from mistaken

judgements about what the pertinent higher-order features are (such as what the relevant principle is justifying the behaviour in question). But there is ample reason to not try and eradicate mere conforming, because this would be to choose failure over success: we can only diminish the range of moral ends that are reached if we don't allow mere conforming, with no concomitant benefit; furthermore, mere conformity is an important avenue by which to introduce individuals (such as generations of children) into the established practices of a community.

6. Social Action without Social Attitudes

It is a prominent position in social philosophy that there is a special kind of attitude on the part of agents when these agents are involved in social action. Variously this attitude is called 'shared intentions' or 'we-intentions' or 'joint commitments'. I will discuss these under the umbrella term *social attitude*. Theorists such as Margaret Gilbert, John Searle, Michael Bratman, Seamus Miller and others have investigated the importance of this kind of attitude. There is good reason to suppose that, just as social actions are ones performed not by individuals acting separately but by collections of individuals working together, social attitudes are not specified by sentences of the form 'I intend that such-and-such', but instead by sentences of the form 'we intend that such-and-such'. There is a debate with Gilbert and Searle on the side for and Bratman and Miller on the side against about whether these social attitudes are a distinct and irreducible kind of attitude, or whether they can be reduced to aggregates of individuals' attitudes. That isn't the question that I am addressing in this chapter. Instead, I am dealing with the different issue about whether social action irreducibly involves social attitudes. I will discuss the views of all the above-mentioned philosophers as varieties of the *social attitude theory* (I will shorten that to *attitude theory* in this chapter). Gilbert goes the furthest, and has developed a comprehensive framework where her favoured version of social attitudes—joint commitments—are the glue of the social world. In her view, the powers and responsibilities that result from participating in a society come from joint commitments by the members to participate in the institutions and practices that produce these powers and place these responsibilities on their participants. My treatment will concentrate on the work of Gilbert and Bratman, because they have given their attitude theories the most recent book-length

defences.[1] By treating prominent examples of both a collectivist and an individualist theory I hope to show that my position cuts across that debate. [2]

The view developed in this thesis has the consequence that individuals can participate in genuinely social action without having a conception of themselves as doing so. That consequence follows from the observation that participating in regularities is a paradigmatically social action, and adding to that the result from Chapter 5 that regularities can persist in the face of ignorance or even error about its nature or purpose. This means that the practices and institutions they are a part of need not figure as such in their attitudes, so that social attitudes need not always be in play. Social actions, then, need not be the object of social attitudes, and hence social attitudes aren't an essential feature of social actions. If my arguments succeed, then this would show that theories like those of Gilbert and Bratman have an important lacuna by concentrating only on those social phenomena with corresponding social attitudes.

In what follows I aim to show that the participants to a regularity need to know nothing more than to do such-and-such in some specific type of circumstance (the behavioural profile of conforming, to use the term introduced earlier in the thesis). Since this means that the attitudes of the participants play no essential role in social action, it also means *a fortiori* that social attitudes do not have such a role.

---

[1] Margaret Gilbert, *Joint Commitment: How We Make the Social World* (Oxford: Oxford University Press, 2013); Bratman, *Shared Agency*.

[2] Though I don't defend the view here, I am happy to acknowledge that regularities in the sense developed by Lewis and myself are irreducibly social phenomena and aren't merely aggregates of individual attitudes. This may be a surprise, since the Lewisian picture with its game-theoretic framework is often taken as a prominent example of an individualist theory. I believe this interpretation is too hasty, because of the manner that regularities feature in individual reasoning. When someone knowingly participates in a regularity, they aren't responding merely to an aggregate of individual behaviours, they are also responding to the combined effect of those behaviours, that being the cooperation brought about by the parties conforming to that regularity. The participants have, through common knowledge of conforming, an avenue for reliably and repeatedly coming to a particular outcome, which is an effect that the individual behaviours do not have the power to bring about. Thus, there is an irreducibly social element to regularities, and thus conventions. For comparison, see a related argument on why methodological individualism cannot exhaust what there is to say about social action Bernard Williams, "Making sense of humanity," in *Making Sense of Humanity: And Other Philosophical Papers 1982–1993* (Cambridge: Cambridge University Press, 1995). Perhaps my collectivist view is in tension with Lewis's, but that remains to be shown.

I call the condition of individuals who participate in a regularity without having a conception of themselves as doing so *rational alienation*. It is a form of alienation in so far as people who are so alienated won't see the ends of these regularities as their own ends, meaning they stand in the same relationship to these ends that Rousseau suggested individuals stand in to the ends of a body politic they are alienated from. This alienation is 'rational' because the way in which there fails to be an integration between the agent and the act is that the end of the act doesn't feature in the individual's reasoning. It is the fullest extent to which social action can be subject to higher-order ignorance, yet I will argue that it is a condition that individuals may find themselves in and suggest that it's not only possible but also familiar.

XXV. Regularities are neutral as to their origin

A useful way to highlight the differences between my regularity-centred view and the attitude theories of Gilbert, Bratman, and others is that I stress that regularities are neutral as regard their origin. What I mean by this is that while social attitudes are very often the source of a regularity (as in, the regularity gets established by a joint commitment or shared agency or whatever), a regularity can be the same complex of behaviours whether its source is a social attitude or anything else. In that respect there is no competition between my view and attitude theories: I am happy to accept joint commitments or shared intentions as interesting accounts of how many social phenomena arise. Nonetheless, I am here pursuing an established avenue to indicate that regularities are more basic than social attitudes, by extending Lewis's result in *Convention* that conventions can arise without explicit agreements. In the social attitudes as analysed by Gilbert and Bratman *et al* I see a generalised and sophisticated development of the pre-theoretical notion of 'agreement' Lewis was contrasting with conventions. This would mean that attitude theories may describe a very widespread aetiology for regularities, but not the necessary development, just as how Lewis showed how conventions can arise without agreements. This is not to say that all social phenomena are instances of regularities, but that

the social phenomena include regularities and regularities don't require social attitudes.

One of the many examples of a convention Lewis provides is a regularity among four people sharing a camp regarding how they collect firewood. Each camper heads out on their own in a different direction (let's say the cardinal directions) in order to collect firewood; in this way they cover the most ground with the least duplicated effort. Lewis uses this example to show how conventions can come apart from agreements. In my hands the example goes through a progression of three different versions, all amounting to the same regularity but each step reducing the role of social attitudes.[3] In the first version the social attitudes are the explicit origin of the regularity. In the second version, some of the participants don't share the social attitudes but the regularity persists regardless. In the third version, none of the participants have the relevant social attitudes.

In the first version of the example all four of the campers agree that each should go in a different direction; a sensible arrangement, since this is the most efficient way. They agree that one camper go north, one go east, another south, and the last one west. Having made this agreement, they go on to each regularly play their specified part. Under Gilbert's analysis, each has made the joint commitment 'for the purposes of collecting firewood, we each cover one of the cardinal directions'; under Bratman's analysis, each shares the intention 'we go each in our own direction such that each covers one of the cardinal directions'. Whatever the particular analysis of the attitudes, in the first version these attitudes are the source of the regularity.

In the second version, consider what happens if one of the campers leaves and is replaced by someone who wasn't party to the founding agreement. The newcomer sees that when the old hands go off to collect firewood each heads north, east, and south respectively, recognises the pattern, and without any explicit direction heads west to collect firewood. By doing so the newcomer takes up the part in the regularity that the departed camper had, and the regularity

[3] The first two of these steps are also in Lewis. The third is original to my expansion.

continues undisturbed. But the newcomer is ignorant about the founding agreement, so cannot have any of the relevant social attitudes that constitute it. The point is that even if they were told about the founding agreement, this would be superfluous: the newcomer can appreciate and participate in the regularity without this knowledge. Lewis goes on to make the further point that in this way all of the original campers can be replaced by newcomers until none of the remaining campers are party to the founding agreement, or even know about it.

The attitude theorist may complain that the newcomer must have some form of joint commitment or shared intention, even if they come to it in some way different from the founding agreement. To use an example used by both by Gilbert and Bratman, the fact that two people are taking a walk together is a matter of their having certain attitudes, but they can have those attitudes without explicitly stating them. However, the most that is required of the newcomer is that what they think and feel be consistent with the social attitudes, and this is too weak to help the attitude theorist. To use the terms developed earlier in the thesis, participating in the regularity constitutes a behavioural profile, and there is a multitude of intentional profiles that would suffice to spontaneously and reliably produce that behavioural profile in individuals. The requirement that the participants think and feel something consistent with the social attitude just is the requirement that those are the motivations that lead to them successfully participating. This in turn just is the requirement that they display the right behavioural profile, without any reference to an intentional profile. And, as I've stressed here, there are many different intentional profiles that will suffice, including the possibility (by way of the alternative method model) that these intentional profiles are radically different from the social attitudes that Gilbert and Bratman *et al* emphasise.

Bratman's theory is the more modest one, so I'll make this point for both theories by way of discussing his. In Bratman's view, there are many different attitudes the parties may have that would suffice for the social action to succeed. He discusses, at length, an example to this

effect from Seamus Miller, about two parties building a tunnel that meets halfway, and considers a variety of different motives parties may have which would still result in the tunnel being built.[4] This looks to be an alternative method model regarding individual attitudes. But Bratman's theory still requires too much: his tunnel-builders are still planning their own actions with reference to the attitudes of the other parties. In contrast, the newcomer in Lewis's camper example doesn't do this. For the newcomer there is only the recognition of the regularity (of the campers each heading in a cardinal direction when collecting firewood) as something like a brute fact.[5] The newcomer can fall into the behaviour required by the regularity without having any awareness or consideration of the attitudes of the other campers, and succeed in what they set out to do, and can rely on being able to succeed. This undermines Bratman's analysis, and *a fortiori* Gilbert's as well. It is worth stating again that this unknowing way to play one's part in a regularity isn't incompatible with having the kind of social awareness that attitude theorists point to—mere conformity doesn't compete with knowing conformity—but that instead the point is to show that regularities can survive the absence of this awareness.

Back to the camper example. For the third version, consider the same regularity but this time without the founding agreement. Each camper heads off in a different cardinal direction to collect firewood, without consulting with each other as to which direction they will go. There are various ways this may have happened, by responding to manifest features of their situation. Perhaps this is because those are the sides of the camp that they have staked out for their own, and thus each camper covers the area that is, as it were, behind their part of the camp. Or perhaps one camper happened to head out north, the second sees this and goes the opposite

[4] Bratman, *Shared Agency*: 51-52, 69-70; Miller, *Character and Moral Psychology*: 13.

[5] Searle, in line with his influential treatment, would say that the campers' regularity isn't a brute fact but instead a social fact because it is the product of an interpersonal behaviour. In contrast, Anscombe argues that facts are brute or not as measured against a social backdrop. So, against the backdrop of our example camp, each camper going into a different cardinal direction to collect firewood would be a brute fact just as much as their biological make-up. My sympathies here lie with Anscombe rather than Searle. John R. Searle, *The Construction of Social Reality* (New York, NY: Free Press); G. E. M. Anscombe, "On Brute Facts," *Analysis* 18, no. 3.

direction (south), the third goes off at a right angle (let's say east) and the fourth the remaining cardinal direction. The fact that each goes the direction most removed from the others is itself open to many different explanations: in order to not interfere with each other, in response to them wanting some time apart from each other, or (the reason that makes this the most efficient outcome) in recognition that doing so means they cover the most ground the most efficiently. Each camper may have a different reason, and may not know or care about the reasons of the other parties. And all of their behaviour may simply be the result of happenstance. The point is, once this pattern has developed and is established, what brought it about doesn't matter. A founding agreement is just one possible aetiology amongst others, and need never to have been in place. What is more, any implementation of the regularity would do, whether it involves social attitudes or not.

Taking a step back, Bratman holds that social actions are based on both shared intentions and the persistent interdependence of individual plans; however, regularities as Lewis and I analyse them are a class of cases where this interdependence alone is enough, and shared intentions are not necessary. As for Bratman, so for Gilbert, and for any other attitude theory.

XXVI. Rational Alienation

Above we have discussed how the existence of a regularity is neutral regarding the origin of the regular behaviour. This allows for higher-order ignorance in the cases where the agents don't know the origin of the regularity. The question is, how far can higher-order ignorance go? Here I argue that it is possible to have regularities where someone has no higher-order knowledge at all: they merely do such-and-such when the recurring situation arises. I don't mean that they play their part in the regularity by accident.[6] What I mean to do here is to show how someone

[6] Care is required here. I do at times claim that the regularity may arise by accident. However, it doesn't follow that once the regularity is established it is only an accident that people conform in any particular instance of the regularity. I mean that people can reliably and repeatedly (that is, non-accidentally) conform to a regularity, even if the origin of that regularity was itself an accident. As I stressed in the previous section,

can reliably and spontaneously conform to a regularity even if they don't know what they're doing or why, what I have called 'rational alienation'.

People who are rationally alienated from a regularity they conform to are in one very real sense just cogs in a machine: the purpose of their actions is opaque to them but instead a feature of a larger process, a process where they are just one party out of many. This seems to be a lamentable condition, but I think it is more than possible: I believe many people are in such a position, and that it is a normal condition for children, teenagers, and perhaps unreflective adults in some parts of their lives. I think it is even encouraged for some people: soldiers and functionaries, for instance, for at least some of their distinctive tasks. Here I don't offer a demonstration of that fact—that would require sociological work that falls outside of the scope of this thesis—but I do provide a demonstration of how rational alienation is possible. And once we recognise that this condition is possible, I believe that we can recognise it as a familiar feature of our social world.

I can harness a familiar philosophical example to also illustrate rational alienation: Ted, the man inside the Chinese Room as imagined by John Searle.[7] Ted works within a room equipped with two slots, a large stack of cards with various intricate designs on them, and a large instruction manual. When a sequence of cards arrives through one slot, Ted refers to the manual and arranges a sequence of cards that he then sends through the second slot in response. Searle asks us to imagine this situation, and then to imagine the further fact that the designs on the cards are Chinese characters, the sequences of cards going in and out of the room are sentences in Chinese writing.[8] Searle's point is that even if you can get sensible answers out of this set-

questions about the origin of a regularity are separable from questions about it persisting. As an analogy, consider how, by the astrophysical laws concerning planetary motion, an astronomical body's becoming trapped in a larger one's orbit is a matter of chance (like the way in which wandering planets or comets or large asteroids can float into a larger body's gravitational field) but once it has been trapped the fact that the smaller body orbits the larger one isn't a matter of chance.

[7] John R. Searle, "Minds, brains and programs," *Behavioral and Brain Sciences* 3, no. 3 (1980).

[8] The example is constructed to mimic a contemporary computer: the cards are the text inputs and outputs, Ted stands for the CPU and the manual for the instruction set.

up, it doesn't mean that Ted can understand Chinese. For instance, if you fed in the sentence 'the building is on fire, you must get out', you may receive a string of cards containing the response 'right, I'll follow you' or 'I must stay and man my post' or 'I can't get out, save yourself', or something similar, but Ted will be ignorant of the fire, despite playing a constitutive role in the production of those sentences. Ted doesn't even know that what he is doing is participating in a communicative act. Ted is in my terms rationally alienated from the act of communicating in Chinese.

It is important to note that the conclusion is easily generalised. We can construct versions of the Chinese Room for any codifiable task, and get the same kind of rational alienation in cases where ignorance like Ted's is possible. This includes very familiar tasks where you could be ignorant of the larger set-up you are a part of. Ted could be performing a categorisation task, like a mail sorter. He could be performing quantitative evaluations, like an insurance underwriter. He could be approving or declining applications based on their fitting a set template, like an administrator. For those of us who have worked inside a large bureaucracy, it is very easy to imagine Ted's situation. Insurance underwriters normally know that they work in insurance, and the same goes for the other examples, but it is possible to imagine someone doing the same work without that knowledge. That would amount to rational alienation.

The claim isn't that there is a limit to how distal an end the individual can have awareness of, and past that limit they can only be rationally alienated. Instead it's that there are some of the ends, not necessarily the most distal, that the rationally alienated individual doesn't have an appreciation of themselves as pursuing. It is, I think, commonplace to have a good grip on your most distal end, but to have significant gaps in your understanding of ends more proximate than that. This would result in you knowing what your ultimate goal is, but not having an appreciation for why the action you are engaged in is meant to be the appropriate means to that goal. Consider an example of an occupation that I think is rife with rational alienation, the

military: presumably it's clear to any soldier that their most distal end *qua* soldier is the security of their nation, yet they may not know how the actions of their regiment contribute to that end.

What is a rationally alienated individual's relationship to the end of the regularity? That is, how does Ted relate to the end of communicating in Chinese? He clearly has no intentional relationship with that end, since he doesn't know about it. That means he has no propositional attitudes that explicitly have that end as their object: he doesn't wish for its fulfilment, wonder about its usefulness, resent it, and so on. What Ted may have, and have in spades, are attitudes which implicitly involve understanding Chinese. Ted stands in a number of actual relationships with the end of understanding Chinese, and by considering these relationships he then implicitly is also considering the end of understanding Chinese. For instance, Ted may notice that the card with '吗' on it appears relatively often, especially at the end of strings of cards. This card represents the question particle in (simplified) Chinese script, which frequently occurs at the end of a sentence.[9] Accordingly, all the attitudes Ted has towards the card '吗' could be cashed out in various ways: as pertaining to the character '吗', the particle '吗' the grammar of questions in written Chinese, word frequency in written Chinese, and so on. But Ted of course doesn't intentionally refer to these things. His position is like that of a baker who frequently considers the effects on dough of kneading and thereby actually but not explicitly refers to the chemistry of gluten.[10] So, Ted's standing towards the end of communicating in Chinese can be of actual import to Ted without him knowing this, just as the chemistry of gluten is of actual import to a baker whether they know anything about it or not. In particular,

[9] This is pronounced 'ma' in Mandarin (in the neutral tone). The awkward way of referring to Chinese linguistic tokens is because there isn't a single language 'Chinese' but instead a broad range of Sinitic languages that share the same script (in simplified and traditional forms), such that communities can share a lot of their written corpus but not be able to communicate verbally. So I treat the example as if there were no spoken form but only a language 'written Chinese' (as there isn't speech in the Chinese Room). The Chinese situation is especially complicated, even by the standards of as difficult a topic as the ontology of language.

[10] Continuing an example used in Chapter 5.

the facts that settle whether a rationally alienated individual like Ted should participate in a regularity or whether he benefits from doing so are independent of whether he knows about them or can knowingly integrate them into his own plans. This is anathema to both streams of attitude theories about social action, exemplified here by the collectivist 'joint commitment' theory of Gilbert and the individualist 'shared intention' theory of Bratman.

XXVII. How does rational alienation persist?

The reader may complain that individuals who are rationally alienated are unlikely to continue conforming to it because it would appear like a purposeless burden. The answer is to point out ways that a population may persist with a regularity that they are rationally alienated from. There are four avenues for this I would like to highlight, and there may be further ones as well. Any one of these avenues would be a sufficient explanation of why individuals may persist with a regularity they are rationally alienated from, and it is likely that an admixture of them would be in play in any one community. Because all that matters for the maintenance of a regularity is mere conforming, which of these avenues (or some other one) happens to describe any particular individual is of no real importance. This I take to be a major contributor to the robustness of regularities, even in the face of rational alienation.

Firstly, the participants may be content to leave well enough alone. It is easy not to appreciate how innocuous something that is the subject of common knowledge usually is. Common knowledge is a feature of your community, and is reaffirmed by each instance where you and your fellows conform to the expectations captured therein. By this token, common knowledge is secure partly because nobody has a salient reason to question it in the face of this constant reaffirmation. This is in striking contrast to the quite common situation in our day-do-day affairs where we are subject to widespread ignorance and have to make decisions under uncertainty. But a grasp of the behavioural profiles captured in common knowledge isn't uncertain—something being common knowledge just means that it's something you know,

expect everybody else to know, expect everybody else to expect you to know, and so on. Accordingly, the fact that there is this unanalysed gap in understanding (regarding higher-order knowledge about the regularity) may not be a point of concern, because the security of common knowledge is itself a respite, being a fixed point in a sea of uncertainty.[11]

The second avenue is by way of the interdependence of individual plans.[12] The fact that we depend on each to accomplish what we ourselves mean to do is something someone can be intimately aware of even in cases where they've lost touch with why they are trying to do these things. The ultimate purpose of an action may be obscure, but that your fellows depend on you to do it nonetheless manifest. While we may not have a grasp on what the ultimate purpose of the other parties' actions are, common knowledge furnishes us with the behavioural profiles of their part in the regularity, so we don't have ignorance about the first-order successes they are after. Similarly, one cannot help but be aware of the extent to which one's own plans depend on the actions of others. So, if there's any end of yours that you can recognise your reliance on others to accomplish, that added to the awareness of your interdependence gives you a reason to facilitate what you understand to be a larger social venture.[13] It would require a mischievous mix of trenchant criticism (of the gap in your understanding of the ultimate aims of the actions) and thoughtless ignorance (or your role in an interdependent social situation) for someone to doubt their reasons for facilitating other people's plans but not question that other people have the respective reasons to facilitate yours. Unfortunately, there are individuals tempted by this kind of mischief. But living in a social world equips you with a rich understanding of what this mischief consists of, and why it is to be avoided. Consider the commonplace and well-realised

[11] Cf. Bernard Williams on skepticism and moral certainty in Bernard Arthur Owen Williams, *Ethics and the Limits of Philosophy*, vol. 83 (Harvard University Press). Ch. 9.
[12] The following takes on board Bratman's discussion of the same in Bratman, *Shared Agency*.
[13] This reason would be a sufficient one if the burdens placed on you by the regularities isn't especially heavy, which it won't be under the specification of benign regularities I have developed in this thesis. Even if this isn't a sufficient reason from the perspective of the agent due to their ignorance of the benefits of the regularity, the mere interdependence is a reason in its own right, which is what I'm trying to establish here.

personality profile of a 'jerk'.[14] Most people recognise that not to play your part in the interdependence of individual actions would be being a jerk, and take this to be a sufficient reason not to do so—and rightly so. For our purposes, it's important to note that this is effective action-guidance in favour of compliance that makes no reference to the justificatory ends that the individuals may be rationally alienated from.

The third avenue is that individuals may think that there are other individuals who know why we do these things, even if they themselves don't. Each participant may think that (enough of) the others know what is going on. The result is a phenomenon usually called *pluralistic ignorance*: a widespread misunderstanding of the social environment where you think that other people are in a different cognitive condition than you are when they are in fact in the same condition, including the presumption that the other people are different from them.[15] So (now taking the perspective of rationally alienated individuals) you may conform to the regularity with the expectation that some other party is providing knowing direction (perhaps you have a particular individual in mind, perhaps not), even though it turns out that the other parties are similarly expecting someone else to offer such direction. In this way everybody's expectation that somebody is a knowing conformer is systematically frustrated, but it doesn't matter because mere conforming is enough.

A fourth avenue is the possibility of pure confabulation. In this case the individuals conform to the regularity, and have a conception of themselves as pursuing some intelligible end to which this regularity is an effective means, but they are mistaken. There isn't a general analysis available of how such instances of confabulation would work, just because it is likely to be *ad hoc*, but there are familiar examples. Here is one: you may conform because you think that the behaviours indicated by conforming are the only good options available, as in Tyler

[14] See Eric Schwitzgebel, "A theory of jerks," *Aeon*(2014), https://aeon.co/essays/so-you-re-surrounded-by-idiots-guess-who-the-real-jerk-is.
[15] See Bicchieri, *The Grammar of Society*: 186-96.

Burge's example of a society that speaks what it takes to be the one natural language.[16]

XXVIII. How widespread could rational alienation be?

Thus far the focus has been on how individuals can play their part within a regularity despite not knowing that they are working towards the regularity's ends, because knowing what to do is separable from knowing what it is you are doing. But this considers only the acts of particular individuals, single threads in a tapestry of social action. The question remains of what the extent of rational alienation could be within a community. There are two dimensions in play: firstly, how many individuals party to a given regularity is rationally alienated (the *popular extent*); and secondly, whether there is a change over time of what proportion of parties are rationally alienated (the *temporal extent*). Instances of rational alienation are easiest to understand when arise when either the popular or the temporal extent isn't total. The first way is where the popular extent of rational alienation isn't total, but where there are some knowing participants who can offer guidance to the unknowing ones. An instance is the familiar example of parents guiding children to ends that the children don't appreciate.[17] The second way is when the temporal extent isn't total, but where a regularity becomes established in a scenario in which the parties (or some of them at least) are knowing participants; suppose that the first-order action guidance gets transmitted without the higher-order knowledge, leaving the later iterations of participants rationally alienated. This is the kind of scenario described by Lewis's campers example.

I will argue that as a limiting case a regularity could display rational alienation to both a total popular and a total temporal extent. My strategy is to first indicate a kind of scenario where the popular extent of rational alienation would be of a total, and then how such a situation could arise without there once having been knowing participants, and thus that it

[16] As discussed in Chapters 1 and 2. See also Burge, "On knowledge and convention."
[17] This example features in Chapter 3.

would also be of total temporal extent. The greater the extent of rational alienation, the less the participation is the result of the foresight of the participants. But even if it happens despite the ignorance of the participants, such a regularity would still be a genuine social good and demand conformity, even though none of its benefactors could explain why this is.

It is easiest to imagine individuals starting to participate in a regularity while lacking any higher-order knowledge of it if they slot into an arrangement that is set up or overseen by someone who does have the relevant higher-order knowledge. This is a familiar state of affairs since that is the condition of children and adolescents as they become socialised into the many regularities of their communities. Often we expect youths to conform first, understand second. So, under my analysis this would mean that youths are rationally alienated for the period where they participate in the regularities that constitute their schooling—conscientiously go to class, do their homework, and study for tests—but haven't yet themselves comes to understand the purpose and value of their education. They participate in their schooling under direction from their parents, teachers, and the wider community, and the institution of schooling would collapse without this direction. In this case, the knowledge of the justification is in effect distributed across the participants, such that there is a core group directors who have that knowledge, while some do without it and depend on the direction they receive.

The above kind of social arrangement is familiar and much-discussed in the philosophical literature by way of treatments of the service conception of authority, as defended by Joseph Raz.[18] In this conception, the purpose of authority is to serve as a supplement to individual decision-making, offering direction to choice-worthy outcomes. Just as under the service conception the subject of authority doesn't need to make the decisions that leads to a choice-worthy outcome, just so the participants to the regularities don't need to appreciate in the

[18] Joseph Raz, *The Morality of Freedom*, vol. 37 (Oxford University Press); Sevel, "The Constitution of Authority."

relevant higher-order features of the regularity to participate and benefit from them.[19] As the development of the service conception has shown, we needn't think of this kind of relationship between knowing directors and unknowing participants as something limited to the case in which the latter are incompetent in some respect (as youths are normally taken to be). Anybody subject to an authority could successfully participate in the regularity by following the commands of the authority. This prominently includes soldiers and functionaries, as I mentioned earlier. However, in the usual treatment of the service conception, it is incompatible with a total popular extent of rational alienation, since there must be knowing directors whose guidance is in the service of the (possibly unknowing) subjects. But under the mere vs knowing conformity model I have developed in this thesis, the role of knowing participants is dispensable.

The benefit of having a knowing participant to a regularity is that they are able to initiate and oversee the running of the regularity. For instance, in the schooling case, it is the parents and the teachers (and perhaps the state authority) that direct the children to attend school, and this is what starts and propagates the regularity of schooling in that community. The knowing party lets the participants know what to do—what the behavioural profile of conforming is. But we have already seen two points that make a knowing participant unnecessary. Firstly, mere conforming suffices for the persistence of the regularity, meaning that the participants having guidance available is separable from whether the source of that guidance is a knowing participant. Secondly, common knowledge of a regularity is an effective mechanism for getting those subject to a regularity to know what participation involves. So, once a regularity has become established well enough to become the subject of common knowledge, the direction of a knowing participant is superfluous. In the limiting case where everybody are mere conformers, this means nobody has higher-order knowledge but the regularity persists

[19] The links between my limited conventionalism and the service conception is also discussed in Chapter 3.

nonetheless, with the same ends and means as it would have had under knowing direction. Thus, once we have a regularity that is kept in place by way of common knowledge or some other mechanism, we may have total popular rational alienation.

Now on to total temporal rational alienation. As noted earlier, the reader may worry that even if we grant the argument for the possibility of total popular rational alienation, the most plausible way that such a situation may arise is by way of the regularity starting with knowing direction, but where the direction withers away over time as the regularity persists by way of mere conforming, with fewer and fewer participants taking the extra step to knowingly conforming.

Here is a suggestion of how to establish the possibility of total temporal rational alienation. The above argument for the possibility of total popular rational alienation made no reference to a vital role played by knowing participants. Instead, it only referred to knowing direction as something that can be replaced. Could a regularity arise without any knowing direction, so that right from the beginning there have only been mere conformers? It is of course conceivable that by brute happenstance a regular behaviour arises which then becomes cemented into a fully-realised regularity. It isn't contradictory to assert that there is a situation where a group of individuals all in the course of their usual affairs happened to stumble on a way in which their disconnected actions happen to coincide into useful cooperation without any of the individuals appreciating this. This would be something like a social 'swampman' case, an analogue to the thought experiment where some matter through astonishingly unlikely happenstance coalesces into a complete duplicate of an individual, down to their memories, dispositions, and perhaps even occurrent thoughts.[20] The social analogue isn't quite as far-fetched, simply by way of there being fewer things that would need to go right by accident.

[20] Donald Davidson, "Knowing one's own mind," *Proceedings and Addresses of the American Philosophical Association* 60, no. 3 (1987).

One clear avenue for establishing the possibility of total temporal rational alienation is to show how there are kinds of social situations where complexes of cooperative behaviours can arise spontaneously without any individual necessarily being aware of it. The thought is that these influences make it likely that people will adopt those behaviours, even if they have no cognitive access to the ends of the regularity that is constituted by their behaving that way. We need to explain how as a limiting case there could be regularities without ever having a knowing participant involved. I now go on to do so.

XXIX. Opaque influences on the rationally alienated

An *opaque influence* is some feature of a situation that makes a difference to the behaviour of an individual but which the individual won't know about just because of this influence. So, it is possible to be affected by an opaque influence but not know that you are. The kind of presuppositions that are uncontroversially built into many social models, such as that individuals pursue what they take to be in their interest, are examples of what I take to be transparent influences.[21] In the context of the current chapter, the kind of social attitudes we find in Gilbert and Bratman *et al* would be transparent, because to be influenced by them means that you are aware of them. This generalises to any interpretation of action where individuals engage in means-ends reasoning. In contrast, to investigate opaque influences is to investigate ways in which individuals' behaviour can be responsive to systematic features of social situations, and even they come to adopt regularities in the face of these situations, without having a conception of themselves as doing so. Here I offer three distinct models of how opaque influences can work on individuals.

The first model tries to salvage a means-end reasoning interpretation even when denying

[21] Some of the other features of the models, such as the assumption of sequential rationality, are controversial often just because they don't seem to be transparent to individuals in the actual situations that these are models for.

that individuals have a conception of themselves as manifesting these behaviours in pursuit of those ends. It models opaque influences as reasoning about ends that work just the way other reasoning about ends does, but without that reasoning coming to the conscious attention of the individual that they are doing so. Call this the *unconscious ends* model. This kind of interpretation is quite common, and prominent examples of it includes much of the work of Sigmund Freud, the 'revealed preferences' model in economics and cognate fields, and explanations of some agents' behaviour in terms of their putative (and often unacknowledged) ideological commitments, and can also be seen in more everyday contexts by way of such locutions as 'this is why they do it, but they won't admit so to themselves'. Many of the uses of such models are deeply controversial, or worse.[22] But we should note that even if these models turn out not to explain the phenomena they are applied to, the cogency and development of an unconscious end model is itself not in dispute. Presumably, it is the cogency and putative explanatory power that makes these models appealing enough to appear and reappear despite their uncertain success as genuine explanations.

The second model gives up on a means-end reasoning interpretation and instead tries to understand opaque influences as a constraint on the options among which individuals can select. It proposes that some options aren't genuinely available to individuals because of the opaque influence (in a sense to be cashed out by an analysis of this particular influence). Call this the *option limitation* model. This may lead to individuals choosing courses of action under some description, but not realising that selecting among the options under the constraint of the opaque influence leads to their behaviour falling under a different description, that of a regularity they are rationally alienated from. It finds expression in, for instance, the live vs dead

[22] There is at best only uncertain empirical support for so-called 'Freudian slips'. The revealed preferences model, while frequently appealed to, is false since it entails behaviourism (by way of positing that mental phenomena are exhaustively described by behaviour) and behaviourism as a theory of mental phenomena is false. That someone's response arises from unacknowledged ideological commitments isn't the kind of thing that could be demonstrated, and is often controversial just because it would be too convenient an explanation. This isn't the venue to pursue these questions any further.

option contrast developed first by William James,[23] which distinguishes between options that are merely possible from those options that are not only possible, but the individual in question has intelligible reasons to pursue. In James's example the opaque influence is the kind of intellectual and motivation background which brings some options to the fore and pushes others back. Another example is from Bernard Williams when he compares ways in which our perspective on distant cultures (distant either geographically or in time) is unlike our perspective on the options within our own culture.[24] In Williams's example the opaque influence is the very different social institutions and established ways of acting which make some options much more viable to pursue than others. It is also implicit in theories where the norms of a society are seen as the result of an interaction between universal demands and the contingencies the society find itself in, such as the theories of David Copp and David Wong.[25] In these kinds of theories it is the resultant norms that are transparent to the members of that society, but the universal demands or the relevant contingencies may very well be opaque influences. And finally, various conventionalist theories also count as option limitation models, such as the appeal to deep conventions proposed by Andrei Marmor,[26] and the limited convention model developed in this thesis. In these theories, the fact that some options are arbitrarily selected and others disregarded by society-wide mechanism is an opaque influence.

The third model exploits the fact that we have some knowledge of the rates at which people display various kinds of behaviour, and that the differences in these rates admit of some kind of systematic explanation, or at least a statistical model. Call this the *differential rates* model. Often part of the explanation why some behaviours are displayed more in certain populations will be factors that people who are under their effect won't automatically know

[23] William James, *The Will to Believe* (Longmans, Green and Co., 1897).
[24] Williams, "Making sense of humanity."; "The Truth in Relativism."
[25] Wong, *Natural Moralities*; "Pluralistic Relativism."; Copp, *Morality, Normativity, and Society*.
[26] Marmor, *Social Conventions*.

about. These would be opaque influences. There is a large literature which provides formal models for dynamic social systems in terms of the differential adoption of various behavioural profiles, such as (to take three examples) the work by theorists like Christina Bicchieri drawing from social epistemology,[27] Robert Sugden drawing from economics,[28] and that of Brian Skyrms drawing from evolutionary game theory[29]. As in the option limitation model, equilibria often feature prominently in differential rate model explanations, though in this case it isn't that the equilibrium property is what highlights some option as especially appropriate to the individual agents, but instead that the features that work upon the individuals lead to predictable equilibria that are exploited as explanations of why we find ourselves in the kinds of situations that we do. For a lot of the work done in this literature (or overlapping literatures) it simply is never pertinent to ask whether individuals are aware of these various opaque influences. Sometimes these models are developed exactly to offer an alternative to explanations that require individuals to be conscious of all the relevant effects of what they're doing.[30]

All three of these three models may be in play in any one population. My own view is that we should look primarily to the second and third of these options, the option limitation and the differential rates models, for the brunt of our explanation of how regularities may arise that the individuals are rationally alienated from. I don't offer an argument that no actual phenomena answers to the unconscious ends model. I will raise the worry that the introduction of unconscious ends is immodest—it requires an explanation both of why the pursuit of that end is efficacious, and why this efficacious pursuit of a desired end (and often the desire itself) is nonetheless hidden from the agent. In contrast, the other two models content themselves with

[27] Bicchieri, *The Grammar of Society*.
[28] Robert Sugden, *The Economics of Rights, Co-Operation, and Welfare* (Oxford: Oxford University Press).
[29] Brian Skyrms, *Evolution of the Social Contract* (Cambridge University Press); *Signals*; *The Stag Hunt and the Evolution of Social Structure* (Cambridge University Press).
[30] As Skyrms does explicitly in, for instance. He goes as far as to posit a 'Darwinian veil of ignorance' where individuals are unwittingly playing their part in improving the fitness of their species. *Signals*: Ch. 1.

explanations of how the opaque influences are efficacious. Furthermore, very often the ascription of unconscious ends is to give a description of an agent that they are likely to object to or which the ascriber means as an objection. These factors incline me to a defeasible preference against the unconscious ends model when other explanations are available.

An instructive example of a theory of social action where the option limitation and differential rate models of opaque influences play a role is the work of Ruth Millikan.[31] Consider, for example, her theory of rule-following by way with an analogy to the mating behaviour of hoverflies.[32] Male hoverflies fly in such a way to intercept zig-zagging dark dots in their field of vision, and if these dots turn out to be female hoverflies the males mate with them. This is an example of differential rates in action because while by no means all zig-zagging dark dots are female hoverflies, a higher proportion of them are than any other item the male hoverfly will have perceptual access to (stationary dots, say). This response to a visual cue thus leads to a differentially higher rate of coming upon female hoverflies. And this is also an example of option limitation, in that the behaviour will persist only in those cases where it leads to enough male hoverflies mating with female ones, and the options where this doesn't happen aren't genuinely available because they lead to the absence of hoverflies. So, the rate of male and female hoverflies mating is an opaque influence for conformity to the rule on the differential rates model, and the fact that only if there are any hoverflies at all if this kind of rule is conformed to is an opaque influence on the option-limitation model. Thus, in Millikan's analysis there is both a dimension where there are factors on the selection that lead to a gradual range of outcomes (the rate at which male hoverflies meet female ones) as well as a dimension where instead there are discrete and qualitative different outcomes (whether the male hoverfly behaviour survives or not). And as for Millikan, so for many other theorists as well.

[31] E.g. Ruth Garrett Millikan, *Language: A Biological Model* (Oxford: Oxford University Press); "Truth, rules, hoverflies."
[32] "Truth, rules, hoverflies."

XXX. Conclusion

This chapter set out to show the possibility that individuals may participate in a social order without understanding themselves as doing so. I take this to be a limiting case to the extent of higher-order ignorance among individuals about the regularities that make up much of that social order. This is the phenomenon that I have called ‘rational alienation’, since by not having cognitive access to the ends of these regularities as the ends that they are pursuing, the agents fail to be integrated with their actions. Despite this being a rather extreme condition, I believe it is a familiar and relatively common one. It is also a blind spot for philosophical theories which takes agents’ self-recognition as participants in a social action as their foundation, theories of like those of Gilbert and Bratman ;the differences between their views (broadly, collectivist vs individualist) don’t change that both kinds of theories depend on social attitudes to explain social actions, but the possibility of mere conformity is also the possibility of someone succeeding in social action despite not having an attitude, or the right kind of attitude, towards that attitude. I have also defended the possibility that it isn’t just isolated individuals who can be so rationally alienated, but that it is possible that there are regularities where no current or past participant have had higher-order knowledge of the ends of that regularity. Finally, I supported that conclusion by a survey of ways in which there can arise a social order by way of opaque influences, meaning that individuals can through these influences arrive at a social order that they do not necessarily have a full epistemic grasp on.

Conclusion

In this concluding chapter, I provide a short sketch of how I take limited conventions to make a difference to moral guidance, and how much of that picture carries over to action-guidance in general. Then I'll give an overview of the work done in this thesis. Finally, I will mark out some future directions for work on limited conventions.

The view of moral guidance I have presented here is of general principles being supplemented by limited conventions. These provide determinate action-guidance in situations subject to the strategic underdetermination problem (SUP). The SUP arises when you need to predict how your fellows will act in order for you to know how you should act (meaning that it the situation is strategic), but the principles which are meant to specify your and their behaviour is underdetermining. Limited conventions address this because they are social regularities that arise from people acting from a structure of shared expectations about how to act. These expectations guide them to coordinate towards a benign outcome, an outcome which is not determinately worse than the others available to you and your fellows. Without these expectations, your uncertainty about what your fellows will do bleeds over into uncertainty about what you should do, for fear of working at cross-purposes with those around you. With the expectations, it is socially settled how you and your fellows should act, and that means you can coordinate towards a benign outcome.

This view of moral guidance amounts to an aggregation of particular conventions for specific repeating situations, that specify what to do where the principles don't say everything. These particular and arbitrary arrangements are kept in place in a community by way of the general expectation of conformity among its members. This doesn't necessarily mean that the members of the community think of themselves as conforming to a convention—they may only think of the principles underlying it and not realise that different responses are possible, or they

may not appreciate the deeper import of conforming. What matters is that so long as there is general conformity to the conventions, the benign outcome in question will arise. Furthermore, every individual is subject to a norm to conform to the convention, because conforming to the convention fulfils a range of nested purposes: it secures the community's cooperation, it works towards the conventionally-selected outcome, and it is a way of conforming to the principles in question. If a convention is established, all someone needs to do to accomplish all of these nested ends as best as they are able is to conform to that convention.

The abovementioned account is general for any kind of action-guidance, and is required for there to be determinate action-guidance in cases where whatever the relevant principles of practical reasoning are in that case (prudential, political, epistemic, and so on), they fail to uniquely specify a response in a given strategic situation. I've discussed the moral case as the most interesting and contested use for conventions, because it isn't seriously doubted that conventions can play a role in practical reasoning more generally. But there are still contributions that my analysis can make in such cases: in particular, highlighting how conventions can be part of a telescoping series of nested purposes, such that they achieve more than just securing one instance of cooperation, but be a constitutive and indispensable part of a larger practice. The above goes beyond conventions, and applies also to the more general notion of a regularity. Conventions arise from the arbitrary selection of one regularity out of a range of different options, but regularities can also be selected when there isn't any underdetermination. The epistemic points I make about regularities selected by a convention and the observations about the ways individuals draw action-guidance from them carry over to any repeated activity in a recurring situation, such as how bakers knead bread in order for it to rise.

XXXI. An overview of the work done in this thesis

The venue for limited conventions is strategic underdetermination problem (SUP) cases. In

Chapter 1 I showed that every settled response to an SUP case is a limited convention, because a settled response leads to a structure of expectations shared across a community about how everybody will act in the given situation. This picks out one particular outcome, that results from the participants each conforming to the expected courses of action. If a benign outcome is picked out, meaning that the outcome is not determinately worse by the lights of the principles than another outcome available to the individuals in question, the result is a Lewisian convention. Since the range of benign outcomes is limited by the principles in play, it is a limited convention. Not conforming to a limited convention harms your fellows by frustrating their ability to reach their desired outcomes without any commensurate benefit, and thus is morally wrong. This wrong is similar to the wrong of lying.

How do such conventions feature in our moral practice? My answer in Chapter 2 is that doing what the convention recommends also has the effect of conforming to a normative principle. I articulated this idea in terms of there being a telescoping series of nested purposes that the relevant action must satisfy. The most proximate purpose is to conform to the convention, more distally is the purpose of playing your part in a specific regularity, more distally still is bringing about a particular outcome, and the most distal is the purpose of conforming to the regularities which make the outcome in question a benign outcome. Individuals can engage with this telescoping series in at least two different ways: a determinative order, where the principles come first and constrain the allowable responses; and an epistemic order, where we often first learn a particular response because it is determinate and available to us through common knowledge, and only later the principle which motivates that response in the given situation. There can also be multiple series of nested purposes relevant to a particular action. While a particular action may be especially pertinent given a particular purpose, it isn't obvious that any of the available descriptions of an action are privileged over another *simpliciter*. I use this framework of nested purposes to counter the

claims of Brennan, Eriksson, Goodin, and Southwood that moral guidance is necessarily practice independent. The framework amounts to a demonstration that conforming to a contingent social practice can also serve as the appropriate way to discharge a moral obligation, and failing to conform to the practice is to fail to conform to the underlying purpose.

This leads to applications of limited conventions. The first one I discuss is how conventions are what makes at least some commands morally obligatory. The thought is that there are social positions someone can occupy such that if that individual issues a command, there is a general expectation throughout a particular community that its members will comply. In the cases where the command would direct the community to a benign outcome, then the general expectation of compliance suffices to establish a limited convention. That in turn means that individual non-conformity to the command would be a moral wrong, because frustrating the reasonable expectations captured in limited conventions is a moral wrong. This is a command-by-command account of conventional authority, providing a criterion for when the command in question is genuinely authoritative: when the command is issued by someone in the appropriate social position and the result would be a benign outcome, the command is authoritative; if not, it would require some other justification to carry weight. I also provide an individual-by-individual account for when someone's commands are genuinely authoritative, by way of the analytic device of a benign arbiter—someone with the appropriate social standing such that compliance to their commands is generally expected, and who commands only things that result in benign outcomes. This reframes the criterion for when conventional authority is justified in terms of whom we should obey: to the extent that someone approximates a benign arbiter, their commands carry genuine authority, otherwise some other justification would be needed. These criteria are instances of what Joseph Raz calls the 'service conception' of authority, which holds that authority is valuable because it helps those subject to it to attain things they value. Using my account, I indicate how parental authority is

an example of conventional authority, since the commands of the parents bring about regularities in the behaviour of their children that benefit the children. In addition, I show how we can have diverging moral commitments within the same community, where the commands of overlapping authorities give rise to nested conventions. That is, different commands with conventional authority in narrower domains (such as the arrangements of individual households) also count as complying to commands in broader domains that also have conventional authority (such as the requirements of what education children in the community must receive).

The above describes how conventions make a difference to what we should do on a case-by-case basis. What then is the overall effect on our normative frameworks? How do our background evaluative practices look if conventions are a ubiquitous feature of moral guidance? In Chapter 4 I discuss this in response to the relativism of David Velleman. Velleman believes that we can't make cross-cultural moral evaluations, because to evaluate an action is to evaluate it as a token of such-and-such action-type, and on his view action-types are culture-specific, and so moral evaluations will be culture-specific as well. I granted Velleman his premises, but offered an account where we can match action-types across cultural boundaries by their evaluative point—the reason why we care how someone handles the action-type in general. For instance, the feudal Japanese action-type '*yū*' and the medieval Italian action-type '*ardimento*' match up with our contemporary Anglophone action-type 'courage', in that all of them are about regulating your response to danger. I concentrated on the action-types, and eventually also the trait-types, that correspond to the virtues and vices because they are the best developed evaluative loaded frameworks of action-types available pre-theoretically—I call these the 'v-types'. I propose that we can accommodate both what is universal and what is particular about them by seeing a society's overall scheme of v-terms as the product of a limited convention. The underlying principles in this case would be the

evaluative points. It is the evaluative points that are universal, while the various ways they can be realised are particular. Such evaluative frameworks also call for a thorough description of how an individual's thoughts, motives, perceptions, and so on, what I call their 'intentional profile', relates to their behaving in some particular way, their 'behavioural profile'. I indicate how there are functional definitions of intentional profiles in terms of behavioural profiles: the intentional profile of a v-type are those psychological features that reliably and spontaneously lead to them displaying the given behavioural profile. Similarly, there are also functional definitions of trait-types in relation to action-types: a virtuous or vicious trait is one that reliably and spontaneously leads to an individual displaying the intentional profile of the respective v-term. In this way, we have a chain of functional definitions linking all the objects of v-terms: the target behaviours are produced by the relevant psychological features, which in turn are produced by the respective character trait.

The analytic tool of behavioural and intentional profiles are of general application, and are especially useful for characterising the relationship individuals have towards the conventions they are expected to conform to. To that end, in Chapter 5 I distinguish two different ways someone can participate in a regularity (conventional or not): knowing conformity, where they display the right intentional and behavioural profiles by conforming to the regularity out of a recognition of the purpose it serves; and mere conformity, where they behave as specified without such a recognition, or based on mistaken judgements about the regularity. I provide the alternative method model for how mere conformity is possible, even commonplace: the recommended action has two different chains of nested purposes, of which one chain features the reasons why that action is the justified one, but the mere conformer instead knows the action only under the description of some different chain which doesn't capture the justification. I identify the ways in which mere conformity is second-best, and the harms that mere conformers are vulnerable to but knowing conformers are protected against. An especially

noteworthy harm are the results of faulty extrapolations, when a mere conformer accepts a mistaken judgement about why they should do something despite being right about what they should do, and on the strength of that mistaken judgement they also do further things which follow from the mistaken judgement and aren't in fact justified. Nonetheless, we should allow mere conformity and prefer it to non-conformity, because it is still a reliable guide to right action without which we would be worse off.

Mere conformity shows that there can be a disconnect between what individuals understand and what they can reliably accomplish. In Chapter 6 I investigate how severe the disconnect could be, and come to the conclusion that it can be total: someone can reliably perform an action that helps secure a purpose without having any conception of themselves as working towards that action. This is what I call rational alienation. This is an especially consequential observation in the face of a range of views that I call 'attitude theories', theories on which certain kinds of action are meant to ineliminably involve a recognition among a group that its members are working towards a shared goal. There is a debate between theorists like Margaret Gilbert who think these social attitudes lead to irreducibly social phenomena, and theorists like Michael Bratman who believe that this amounts to nothing more than aggregations of individual attitudes. But cutting across this dispute, actions arising from conventions are paradigmatic social actions but with a structure of expectations that make rational alienation possible. This means that mere conformers to a convention needn't have any conception of themselves or their fellows as working towards some given purpose, even if they reliably do things that secure that purpose.

I go on to discuss three different models for how rational alienation may arise: an unconscious ends model, where individuals don't realise their own motivations; an option limitation model where features of actions are such that the only courses of action that secure some specific purpose are those that are liable to be reliably followed, irrespective of whether

anyone is aware of these features; and a differential rates model where underlying features of the population lead to some course of action disproportionally being followed and where this preponderance lead to some courses of action crowding out other options, irrespective of anyone being aware of this preponderance. I indicate that the unconscious ends model can't explain equilibrium phenomena in populations. This, along with its immodest psychological commitments, make me recommend a defeasible preference for turning to the option limitation and differential rate models for explaining instances of rational alienation.

XXXII. Further directions for work on limited conventions

The uses I have put limited conventions to in this thesis doesn't exhaust what can be done with them. One prominent further avenue to explore is to apply them to the domain of law. On the occasions when I have made use of examples that involve the law I have emphasised their standing as practical and social arrangements, rather than leveraging a special standing they may have as law. But such a special standing that the law seems to have is part of what needs to be explained. Similarly, I've never discussed anything like the standing to punish non-conformers, and punishment simply doesn't enter my analysis at any point, while it is standardly taken as an important feature of the law. So work remains to be done, linking the general framework I have developed here to conventionalist analysis of the law such as Marmor's. This would amount to identifying some underlying social arrangements that find expression and are made determinate and binding by surface conventions codified by law.

Another avenue that limited conventions can be put to use to is to expand on the ways in which it makes obligations, moral and otherwise, be situated in a particular context. In the thesis I have described mechanisms by which contingent expressions of underlying principles are binding, despite the fact that they may diverge from context to context and that individuals may, in the first instance, relate to the conventions and regularities in play rather than the underlying principles. This goes some way to explaining how moral guidance can be situated,

but there is more to be done here. We would need an extensive treatment of how conventions respond to historical contingencies, and how they may change, wither away, or be renewed. I haven't had the space to do that here. There is a growing literature on the evolution of norms, and I have at times referred to some of the contributors to that literature, which prominently includes Christina Bicchieri, Brian Skyrms, Robert Sugden, and Brennan, Eriksson, Goodin, and Southwood. But their work isn't in general situated in the sense I mean: they describe general mechanisms which allow for social arrangements to come into being and pass away. For it to be an explanation for situated social arrangements, as all actual social arrangements patently are, they need another mechanism on top of the broadly rational-choice-theoretic framework they adopt in order to make determinate the abstract conclusions that arise for their investigation into classes of games. Bicchieri, in her recent work at the Behavioural Ethics Lab and the Social Norms Training and Consulting Group (both at the University of Pennsylvania), and with UNICEF on changing norms regarding female genital mutilation/cutting, are examples of the intersection between the general frameworks and situated social arrangements. I see a similar use for limited conventions.

Finally, while I have avoided discussing what the underlying principles may be like in order to stress that whatever they are, they need limited conventions in order to address SUP cases, I have no intention to remain uncommitted on this point in general. Limited conventions are especially pertinent for views on the underlying principles that allow for massive underdetermination, since such underdetermination is often seen to be an objectionable feature; the work on limited conventions done here shows that they can be workable. There are two avenues which I find especially suggestive. The first is something like a Kantian view, where there is a general overarching framework (in his case, the categorical imperative) that identifies certain classes of actions as required based on how they relate to meaningful action being possible at all. Limited conventions could be a way to take such an overarching framework and

draw from it a concrete and particular system of moral guidance, in line with Hegel's critique of Kant, Rawls's suggestions on this point and work in the Rawlsian spirit like Michael Walzer's 'Liberalism A and Liberalism B' cited in chapter 3, and the recent of work of theorists like Onora O'Neill, Barbara Herman, and Talbot Brewer. The second suggestive approach is something in the broad family of views that identify moral practices as an innate natural feature of humans and human communities. Aquinas has very suggestive but underdeveloped suggestions on this point with his reconciliation between natural and human law in his *Treatise on Law*, and naturalism in the style of Philippa Foot is a programmatic statement of how this may work. From a different tradition, David Wong understands his pluralistic relativism to be leveraging the same kind of concerns. Both these kinds of view have broad appeal, and both are frequently attacked because of the difficulty of drawing determinate guidance from them. Limited conventions could be a way to take on board what is appealing about them, and also to make them practicable.

Appendix A: Functional definitions and the situationist challenge to the virtues

In Chapter 4 I address evaluatively loaded action-types, concentrating on those related to the virtues and vices. I gave an argument that we should extend our treatments of action-types to also cover the corresponding trait-types. I settled on calling the virtue- and vice-related action-types *v-acts* and the corresponding trait-types *v-traits*, using *v-types* to cover both. There I proposed and defended a way of linking action- and trait-types, the functional definition thesis:

> *each v-type includes both instances of actions (v-acts) and character traits (v-traits); v-traits can be defined in terms of v-acts, such that the v-traits are those character traits that in the relevant circumstances lead to their possessor spontaneously performing v-acts.*

This is of special relevance to the so-called *situationist challenge* to virtue ethics. The situationist challenge is the introduction into philosophy, most prominently by Gilbert Harman, of an interpretation of experiments in social psychology meant to show that individuals don't have stable character traits which lead them to reliably act in the relevant way across a range of situations. [1] In John Doris's milder version, individuals still possess limited *local* character traits, where they have stable dispositions to behave in regular ways in tightly-specified situations, but not *global* character traits that hold across situations.[2] In both Harman and Doris's positions (taking these as the canonical statements of the situationist challenge) any theory that gives traits related to the virtues a robust explanatory role is an unviable system,

[1] Gilbert Harman, "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error," *Proceedings of the Aristotelian Society* 99, no. 1999 (1999).
[2] John M. Doris, *Lack of Character* (Cambridge: Cambridge University Press, 2002).

since it is meant to require stable and global traits that these theorists believe we have reason to deny the existence of.[3]

My response is to use the functional definition thesis to show how virtue theories have the resources to bridge the kind of evaluation situationists allow for, with actions taken as primary, and the kind of evaluation they attack, where traits are taken as primary. In the literature there are attempts to respond to the situationist challenge by offering action-centric conceptions of virtue, what Thomas Hurka and Gopal Sreenivasan call the 'occurrent act' account of virtue.[4] These are meant to be revisions of our virtue theory to address their vulnerability to the situationist challenge. What I am doing here is showing that because of the functional definition thesis we don't need to make any modifications to traditional virtue theories in order to take actions as primary. This means that the resources which these modified theories have attempted to acquire are already available to traditional virtue theories.

Here are some indications to show that taking actions as primary is acknowledged as a sufficient response to the situationist challenge. In later work Harman moderates his skepticism about character traits, admitting that some accounts of the virtues escape the situationist challenge, and that the challenge doesn't entail the nonexistence of character traits. Instead, it is meant to make their existence not be obvious—especially for the kind of robust traits required by many prominent virtue theories. He reaffirms that what is at stake is action explanation in terms of traits that goes beyond what can be done by way of action explanation in terms of acts. In his words:

---

[3] Where it matters, I will distinguish between full-blown virtue ethics and so-called 'virtue theories' which gives the virtues robust explanatory roles without endorsing virtue ethics. Examples of virtue theories include Aquinas, "Treatise on Law."; "Quaestiones disputatae de virtutibus."; Kant, *The Metaphysics of Morals*; Peter Geach, *The Virtues* (Cambridge University Press); Robert Merrihew Adams, *A Theory of Virtue: Excellence in Being for the Good* (Clarendon Press, 2006).

[4] Hurka, "Virtuous acts, virtuous dispositions," 75; Gopal Sreenivasan, "The situationist critique of virtue ethics," in *The Cambridge Companion to Virtue Ethics*, ed. Daniel C. Russell (Cambridge: Cambridge University Press, 2013).

> *Virtue or character as a fleeting feature of an act must be distinguished from virtue or character as an enduring characteristic of a person. There is more reason to believe that there are virtuous and vicious acts than to believe that people have virtuous or vicious characters.* [5]

Furthermore, in his treatment of Judith Jarvis Thompson's virtue theory, Harman says that it avoids the situationist challenge because the purported stable traits can be redefined in terms of collections of actions whose existence nobody denies.[6]

The functional definition thesis shows that we can produce informative specification of traits just in terms of some set of target behaviours, meaning that virtue theories have the resources to close the gap between action explanation in terms of actions and in terms of traits. This undercuts the canonical form of the situationist challenge. It also undercuts those responses to the challenge that strip virtue ethics of global traits while retaining an analogue of the v-acts, such as by Maria Merritt and Mark Alfano.[7] What is more, my theory is hospitable to every major virtue theory. So, the situationist challenge addresses no major virtue theory. This diagnosis helps explain why the majority of responses to the situationist challenge has been to dismiss it as missing its target.[8]

Even if the situationist were to grant that acts and traits can be bridged in this way, they may press the point to show that at best to talk about traits in this way would be a failed posit.

---

[5] Gilbert Harman, "Skepticism about Character Traits," *Journal of Ethics* 13, no. 2/3 (2009): 239-40, 41.

[6] "Moral Philosophy Meets Social Psychology," 327-28. A similar concession is made by Doris, *Lack of Character*: 116-17. See also Judith Jarvis Thomson, "The right and the good," *Journal of Philosophy* 94, no. 6 (1997); Snow, *Virtue as Social Intelligence*: 5, 8-11; Hurka, "Virtuous acts, virtuous dispositions," 75; Sreenivasan, "The situationist critique of virtue ethics."

[7] Maria Merritt, "Virtue ethics and situationist personality psychology," *Ethical Theory and Moral Practice* 3, no. 4 (2000); Mark Alfano, "Identifying and Defending the Hard Core of Virtue Ethics," *Journal of Philosophical Research* 38(2013).

[8] Of the many examples of this response, I would like to highlight Sreenivasan's to the effect that the experiments cited are of the wrong kind to draw conclusions about the type of character traits in question. In addition to other worries about the cited studies, like that they don't properly operationalise virtuous behaviour, he stresses that almost all of them are one-off tests, whereas to say anything about an individual's character traits you would need to test that same individual in a variety of situations, e.g. make use of an iterated trial experiment: Gopal Sreenivasan, "Errors about errors: Virtue theory and trait attribution," *Mind* 111, no. 441 (2002); "Character and consistency: Still more errors," *Mind* 117, no. 467 (2008); "The situationist critique of virtue ethics.".

This is because of empirical results where individuals' actions vary widely because of tiny situational factors, in a way that seems to swamp out any personal characteristics. This is meant to show that the traits aren't operative, however we understand them. But this would be to overinterpret the empirical results. What the situationist challenge highlights is that there doesn't seem to be evidence that individual actions cohere across different situations, when understood in terms of their objective construals—that is, their behavioural profiles. However, the functional definition thesis as defended here doesn't identify traits by way of a consistency in behavioural profiles, but instead through consistency of intentional profiles. Any one behavioural profile can be caused by a wide variety of different intentional profiles—this is the old point that the behaviours that match up to mental states are multi-track dispositions. So we have no reason to expect that there is any one behavioural profile—that is, objective construal—that different instances of acting from a virtue would correspond to. When we measure an individual's actions by way of their subjective construals of the situation they are acting in, we have very good evidence for exactly the type of cross-situational consistency the situationist challenge was meant to show doesn't exist.[9] And, of course, subjective construals are a part of the intentional profile of an action and a trait.[10]

When we highlight the fact that consistency in behaviour needs to be understood as going alongside consistency in subjective construal, the results that Harman and Doris and others want us to see as the proof that individuals don't display traits consistently is at most proof that it is easy to come up with situations which individuals don't construe as requiring them to act in such-and-such a manner consistent with possessing a virtue. But even that result is too strong: many people who fail to act consistent with the objective construal nonetheless showed signs

[9] "Character and consistency."; Rachana Kamtekar, "Situationism and virtue ethics on the content of our character," *Ethics* 114, no. 3 (2004); Snow, *Virtue as Social Intelligence*.
[10] Subjective/objective construals and behavioural/intentional profiles are extensively discussed in Ch. 4, §IV.

that they had recognised that the situation may call for the behaviour the experiment coded as possessing the virtue.[11] They would normally cite countervailing reasons for not displaying the looked-for behaviour. Even if their countervailing reasons turn out not to be very compelling (maybe they are, maybe they aren't), the mere attempt to provide an explanation indicates that people are sensitive to the fact that such-and-such behaviour is called for. The observation that many people are too sensitive to countervailing reasons and not sensitive enough to the reasons of virtue is no news at all, and no serious virtue theory is threatened thereby.

The above response hasn't escaped criticism. It has been argued that if the virtue theorists can only save their theory in the face of the situationist challenge if they make the virtues (or vices) out to exist very rarely, then that robs the theory of its appeal: the virtues are meant to be a widespread and everyday evaluative framework, and it would be hard to see how this could be the case if they are almost never to be seen. This is often called the 'rarity thesis' and it is a matter of debate whether, even if true, this is a compelling way to press the point of the situationist challenge.[12] But the functional definition thesis allows us to sidestep this worry. For the rarity thesis to make trouble for the virtue theorist, it needs to be the case that if the virtues aren't commonplace, it isn't plausible that people can widely make evaluations in terms of them. But since we can specify v-traits while referring only to the v-acts, the v-traits can be widely known if the v-acts are. And it isn't contested that the v-traits are widely known. What is at issue isn't that someone can see the virtues in the flesh—though of course v-acts are instantiations of virtue and vice—but that they are able to tell what possessing a virtue would

[11] The objective construals according to which these individuals didn't act consistently with the possession of a virtue have to be extremely narrow, since their behaviour clearly shows many of them to be conflicted about not doing the virtuous thing. See the discussion of this effect in various situationist studies in Snow, *Virtue as Social Intelligence*: 100-16. See also Sreenivasan, "The situationist critique of virtue ethics," 300-03.

[12] For examples of people making this objection: Alfano, "Identifying and Defending the Hard Core of Virtue Ethics."; J. S. Blumenthal-Barby, "Dilemmas for the Rarity Thesis in Virtue Ethics and Virtue Epistemology," *Philosophia* 44, no. 2 (2016). For one response, see Micah Lott, "Situationism, Skill, and the Rarity of Virtue," *Journal of Value Inquiry* 48, no. 3 (2014).

be like. And a specification by way of functional definitions are a perfectly good way to do this. The analysis I have provided is technical, but it is meant to be a technical presentation of something that is an everyday feature of evaluations: linking an individual act with a trait that would suffice to produce it.

So, to summarise how the functional definition thesis addresses the situationist challenge: the situationists argue that there is an empirical refutation of the claim that there are character traits that are operational influences in how individuals behave since situational factors on behaviour seem to swamp out the influence of stable dispositions of character; one response that is invited by the situationists is to move the focus of evaluation from character traits to individual actions; but the functional definition thesis shows that we can define the v-types taking actions as primary without giving up the import of traits. This sharpens our understanding of what would be required to show that virtues and vices aren't in effect in the kinds of studies that the situationists point to, and how they can persist as a pre-theoretical evaluative framework even if pure virtues or vices are very rare.

Bibliography

Adams, Robert Merrihew. *A Theory of Virtue: Excellence in Being for the Good*. Clarendon Press, 2006.

Alfano, Mark. "Identifying and Defending the Hard Core of Virtue Ethics." *Journal of Philosophical Research* 38 (2013): 233-60.

Annas, Julia. *Intelligent Virtue*. Oxford: Oxford University Press, 2011.

Anscombe, G. E. M. *Intention*. Cambridge, MA: Harvard University Press, 1957.

———. "On Brute Facts." *Analysis* 18, no. 3 (1957): 69-72.

Aquinas, Thomas. "Quaestiones Disputatae De Virtutibus."

———. "Treatise on Law." In *Summa Theologica*, I-II QQ 90-108.

Aristotle. "Nicomachean Ethics."

Baumrind, Diana. "Authoritative Parenting Revisited: History and Current Status." In *Authoritative Parenting*, edited by Robert E. Larzelere, Amanda Sheffield Morris and Amanda W. Harrist. Washington, DC: American Psychological Association, 2013.

Bicchieri, Cristina. *The Grammar of Society*. Cambridge: Cambridge University Press, 2005.

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261-325.

Blumenthal-Barby, J. S. "Dilemmas for the Rarity Thesis in Virtue Ethics and Virtue Epistemology." *Philosophia* 44, no. 2 (2016): 395-406.

Braddon-Mitchell, David, and Frank Jackson. *The Philosophy of Mind and Cognition*. 2nd ed. Oxford: Blackwell, 2007.

Bratman, Michael E. *Shared Agency: A Planning Theory of Acting Together*. New York, NY: Oxford University Press, 2014.

Braybrooke, David. "No Rules without Virtues; No Virtues without Rules." *Social Theory and Practice* 17, no. 2 (1991): 139-56.

———. *Utilitarianism: Restorations, Repairs, Renovations*. Toronto: University of Toronto Press, 2004.

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood. *Explaining Norms*. Oxford: Oxford University Press, 2013.

Brewer, Talbot. "Maxims and Virtues." *Philosophical Review* 111, no. 4 (2002): 539-72.

Burge, Tyler. "On Knowledge and Convention." *Philosophical Review* 84, no. 2 (1975): 249-55.

Cook, Megan. "Abortion." In *Te Ara Encyclopedia of New Zealand*, edited by Manatū Taonga Ministry for Culture and Heritage, 2011.

Copp, David. *Morality in a Natural World: Selected Essays in Metaethics*. Cambridge: Cambridge University Press, 2007.

———. *Morality, Normativity, and Society*. Oxford: Oxford University Press, 1995.

Dancy, Jonathan. "In Defense of Thick Concepts." *Midwest Studies in Philosophy* 20, no. 1 (1995): 263-79.

Davidson, Donald. "Knowing One's Own Mind." *Proceedings and Addresses of the American Philosophical Association* 60, no. 3 (1987): 441-58.

Doris, John M. *Lack of Character*. Cambridge: Cambridge University Press, 2002.

Driver, Julia. *Uneasy Virtue*. Cambridge: Cambridge University Press, 2001.

Finnis, John. *Aquinas : Moral, Political, and Legal Theory*. Oxford: Oxford University Press, 1998.

Gallagher, Shaun. "Direct Perception in the Intersubjective Context." *Consciousness and Cognition* 17, no. 2 (6// 2008): 535-43.
Gaus, Gerald. "Review of 'Explaining Norms'." In, *Notre Dame Philosophical Reviews* (2014). http://ndpr.nd.edu/news/explaining-norms/.
Gauthier, David P. *Morals by Agreement*. Oxford: Oxford University Press, 1986.
Geach, Peter. *The Virtues*. Cambridge University Press, 1977.
Gilbert, Margaret. "Agreements, Conventions, and Language." *Synthese* 54, no. 3 (1983 1983): 375-407.
———. "Critical Notice: Gilbert Harman and Judith Jarvis Thomson, Moral Relativism and Moral Objectivity." *Noûs* 33, no. 2 (1999): 295-303.
———. *Joint Commitment: How We Make the Social World*. Oxford: Oxford University Press, 2013.
———. *On Social Facts*. Princeton, NJ: Princeton University Press, 1992.
Goldoni, Marco. "Multilayered Legal Conventionalism and the Normativity of Law." In *The Normative Dimension of Law*, edited by S Bertea and G Pavlakos. 158-76. Oxford: Hart Publishing, 2011.
Harman, Gilbert. "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error." *Proceedings of the Aristotelian Society* 99, no. 1999 (1999): 315-31.
———. "Moral Relativism Defended." *Philosophical Review* 84, no. 1 (1975): 3-22.
———. "Skepticism About Character Traits." *Journal of Ethics* 13, no. 2/3 (2009): 235-42.
Hart, H. L. A. *The Concept of Law*. Oxford: Oxford University Press, 1994.
Herman, Barbara. *Moral Literacy*. Cambridge, MA: Harvard University Press, 2007.
Hurka, Thomas. "Virtuous Acts, Virtuous Dispositions." *Analysis* 66, no. 1 (2006): 69–76.
Hursthouse, Rosalind. "The Central Doctrine of the Mean." In *The Blackwell Guide to Aristotle's Nicomachean Ethics*, edited by Richard Kraut. 96-115. Malden, MA: Blackwell, 2006.
———. "Intention." *Royal Institute of Philosophy Supplement* 46 (2000/003/001 2000): 83-105.
———. *On Virtue Ethics*. Oxford: Oxford University Press, 1999.
Jackson, Frank. *From Metaphysics to Ethics*. Oxford: Oxford University Press, 1998.
Jackson, Frank, and Philip Pettit. "Moral Functionalism and Moral Motivation." *The Philosophical Quarterly* 45, no. 178 (1995): 20-40.
———. "Moral Functionalism, Supervenience and Reductionism." *The Philosophical Quarterly* 46, no. 182 (1996): 82-86.
James, William. *The Will to Believe*. Longmans, Green and Co., 1897.
Kamtekar, Rachana. "Situationism and Virtue Ethics on the Content of Our Character." *Ethics* 114, no. 3 (2004): 458-91.
Kant, Immanuel. *The Metaphysics of Morals*. edited by Mary J. Gregor Cambridge: Cambridge University Press, 1996. 1797.
Lewis, David. *Convention: A Philosophical Study*. Malden, MA: Harvard University Press, 1969.
———. "Languages and Language." *Minnesota Studies in the Philosophy of Science* 7, no. Journal Article (1975): 3-35.
———. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50, no. 3 (1972): 249-58.
———. "Truth in Fiction." *American Philosophical Quarterly* 15, no. 1 (1978): 37-46.
Lott, Micah. "Situationism, Skill, and the Rarity of Virtue." *Journal of Value Inquiry* 48, no. 3 (2014): 387-401.

MacIntyre, Alasdair C. *A Short History of Ethics*. Notre Dame, IN: University of Notre Dame Press, 1966.
Marmor, Andrei. *Social Conventions: From Language to Law*. Princeton, NJ: Princeton University Press, 2009.
Merritt, Maria. "Virtue Ethics and Situationist Personality Psychology." *Ethical Theory and Moral Practice* 3, no. 4 (2000): 365-83.
Miller, Christian. *Character and Moral Psychology*. Oxford: Oxford University Press, 2014.
Miller, Seumas. *Social Action: A Teleological Account*. Cambridge: Cambridge University Press, 2001.
Millikan, Ruth Garrett. "A Difference of Some Consequence between Conventions and Rules." *Topoi* 27, no. 1-2 (2008): 87-99.
———. "Language Conventions Made Simple." *Journal of Philosophy* 95, no. 4 (1998): 161-80.
———. *Language: A Biological Model*. Oxford: Oxford University Press, 2005.
———. "Truth, Rules, Hoverflies, and the Kripke-Wittgenstein Paradox." *Philosophical Review* 99, no. 3 (1990): 323-53.
Morgan, Jeffrey. "Children's Rights and the Parental Authority to Instill a Specific Value System." *Essays in Philosophy* 7, no. 1 (2006).
Murphy, Mark C. *Natural Law in Jurisprudence and Politics*. Cambridge: Cambridge University Press, 2006.
Noggle, Robert. "Special Agents: Children's Autonomy and Parental Authority." In *The Moral and Political Status of Children*, edited by David Archard and Colin M. Macleod. 97-117. Oxford: Oxford University Press, 2002.
Nussbaum, Martha C. "Non-Relative Virtues: An Aristotelian Approach." *Midwest Studies In Philosophy* 13, no. 1 (1988): 32-53.
O'Neill, Onora. *Towards Justice and Virtue: A Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press, 1996.
Pettit, Philip. "The Consequentialist Perspective." In *Three Methods of Ethics: A Debate*, edited by Marcia W. Baron, Philip Pettit and Michael Slote. 92-174. Oxford: Blackwell, 1997.
Pettit, Philip, and Geoffrey Brennan. "Restrictive Consequentialism." *Australasian Journal of Philosophy* 64, no. 4 (1986): 438 – 55.
Pippin, Robert B. *Hegel's Practical Philosophy: Rational Agency as Ethical Life*. Cambridge: Cambridge University Press, 2008.
Rachels, James, and Stuart Rachels. *Elements of Moral Philosophy*. 8th ed. New York, NY: McGraw-Hill, 2015.
Raz, Joseph. *The Authority of Law*. 2nd ed. Oxford: Oxford University Press, 2009. doi:10.1093/acprof:oso/9780198253457.001.0001.
———. "Introduction." In *Authority*, edited by Joseph Raz. New York, NY: New York University Press, 1990.
———. *The Morality of Freedom*. Vol. 37: Oxford University Press, 1986.
Russell, Daniel C. *Practical Intelligence and the Virtues*. Oxford: Oxford University Press, 2009.
Schwitzgebel, Eric. "A Theory of Jerks." In, *Aeon* (2014). https://aeon.co/essays/so-you-re-surrounded-by-idiots-guess-who-the-real-jerk-is.
Searle, John R. *The Construction of Social Reality*. New York, NY: Free Press, 1995.
———. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417-57.
Sevel, Michael. "The Constitution of Authority: A Review of Joseph Raz, between Authority and Interpretation: On the Theory of Law and Practical Reason." *Jurisprudence* 5, no. 2 (2014): 430-41.

Shaver, Robert. "Virtues, Utility, and Rules." In *The Cambridge Companion to Adam Smith*. Cambridge: Cambridge University Press, 2006.
Skyrms, Brian. *Evolution of the Social Contract*. Cambridge University Press, 1996.
———. *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press, 2010.
———. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, 2012.
"Sleep-Wake Disorders." In *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 2013.
Slote, Michael A. *Morals from Motives*. Oxford: Oxford University Press, 2001.
Snow, Nancy E. *Virtue as Social Intelligence*. New York, NY: Routledge, 2010.
Southwood, Nicholas. "The Authority of Social Norms." In *New Waves in Metaethics*, edited by Michael Brady. 234-48. Houndmills: Palgrave Macmillan, 2011.
———. "The Moral/Conventional Distinction." *Mind* 120, no. 479 (2011): 761-802.
Southwood, Nicholas, and Lina Eriksson. "Norms and Conventions." *Philosophical Explorations* 14, no. 2 (2011): 195-217.
Spiekermann, Kai. "Review of 'Explaining Norms'." *Economics and Philosophy* 31, no. 1 (2015): 174-81.
Sreenivasan, Gopal. "Character and Consistency: Still More Errors." *Mind* 117, no. 467 (2008): 603-12.
———. "Disunity of Virtue." *Journal of Ethics* 13, no. 2-2 (2009): 195-212.
———. "Errors About Errors: Virtue Theory and Trait Attribution." *Mind* 111, no. 441 (2002): 47-68.
———. "The Situationist Critique of Virtue Ethics." In *The Cambridge Companion to Virtue Ethics*, edited by Daniel C. Russell. 290-314. Cambridge: Cambridge University Press, 2013.
Stern, Robert. "On Hegel's Critique of Kant's Ethics: Beyond the Empty Formalism Objection." In *Hegel's Philosophy of Right: Essays on Ethics, Politics, and Law*, edited by Thom Brooks. 73-100. Malden, MA: Blackwell, 2012.
Sugden, Robert. "Contractarianism and Norms." *Ethics* 100, no. 4 (1990): 768-86.
———. *The Economics of Rights, Co-Operation, and Welfare*. Oxford: Oxford University Press, 1986.
Swanton, Christine. "A Virtue Ethical Account of Right Action." *Ethics* 112, no. 1 (2001): 32-52.
———. *Virtue Ethics: A Pluralistic View*. Oxford: Clarendon Press, 2003.
"Termination of Pregnancy in New Zealand." edited by Best Practice Advocacy Centre New Zealand. Dunedin, 2010.
Thomson, Judith Jarvis. "The Right and the Good." *Journal of Philosophy* 94, no. 6 (1997): 273-98.
van Roojen, Mark. *Metaethics: A Contemporary Introduction*. Oxford: Routledge, 2015.
Velleman, J. David. "Doables." *Philosophical Explorations* 17, no. 1 (2014/01/02 2014): 1-16.
———. *Foundations for Moral Relativism*. 2nd ed. Cambridge: OpenBook Publishers, 2015.
Verbeek, Bruno. "Conventions and Moral Norms: The Legacy of Lewis." *Topoi* 27, no. 1-2 (2008/07/01 2008): 73-86.
Walzer, Michael. "Comment." In *Multiculturalism*, edited by Amy Gutman. 99-104. Princeton, NJ: Princeton University Press, 1994.
Williams, Bernard. "Making Sense of Humanity." In *Making Sense of Humanity: And Other Philosophical Papers 1982–1993*. 79-89. Cambridge: Cambridge University Press, 1995.
———. "The Truth in Relativism." *Proceedings of the Aristotelian Society* 75 (1974): 215-28.

Williams, Bernard Arthur Owen. *Ethics and the Limits of Philosophy*. Vol. 83: Harvard University Press, 1985.
Wong, David B. *Natural Moralities: A Defense of Pluralistic Relativism*. New York: Oxford University Press, 2006. doi:10.1093/0195305396.001.0001.
———. "Pluralistic Relativism." *Midwest Studies In Philosophy* 20, no. 1 (1995): 378-99.
Yablo, Stephen. "Definitions, Consistent and Inconsistent." *Philosophical Studies* 72, no. 2-3 (1993): 147 - 75.
Young, H. Peyton. "The Evolution of Social Norms." *Annual Review of Economics* 7, no. 1 (2015): 359-87.
Zagzebski, Linda. *Divine Motivation Theory*. Cambridge: Cambridge Univeristy Press, 2004.
———. "Exemplarist Virtue Theory." *Metaphilosophy* 41, no. 1 (2010): 41-57.