The University of Southampton

FACULTY OF ARTS & HUMANITIES

Department of Philosophy

Subjective Normativity:
Understanding Reasons and 'Oughts' as Contingent on Desires

by

Elizabeth Ventham

Thesis for the Degree of Doctor of Philosophy

August 2018

Abstract

This thesis argues that what an agent has reason to do, and what an agent ought to do, are contingent on that agent's desires. Unless that agent has some desire that could be satisfied (or that the agent believes could be satisfied) by an action, then that agent has no reason to choose to act in that way, and it is not the case that they ought to act in that way.

I will argue for this subjective account of normative reasons and oughts across four chapters. The first two chapters will defend the desire-based account of reasons. I will explain two positive arguments in Chapter 1, one about capacity for action and one about non-desire-based reasons as different kind of phenomena. For the rest of the chapter and Chapter 2 I will defend the account against three main objections, one that can be attributed to McDowell and two to Parfit. I will also use Chapter 2 to defend 'value subjectivism' – the theory that what's valuable to an agent is contingent on an agent's desires. This will be used to support my arguments for desire-based reasons and oughts.

The second half of my thesis will argue that what we ought to do is based on our desires. Chapter 3 will build on the work done in the previous chapters and demonstrate that my subjective accounts are compatible with a wide range of qualities that we want normative oughts to have. It will also respond to two objections, and argue against a rival account of oughts: that of categorical imperatives. Chapter 4 will then defend my account against its final rival: an account on which there are 'overall oughts' that aren't based on an agent's desires.

Contents

Author's Declaration

UNIVERSITY OF Southampton

Research Thesis: Declaration of Authorship

Print name:	Elizabeth Ventham

Title of thesis:	**Subjective Normativity: Understanding Reasons and 'Oughts' as Contingent on Desires**

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

_____ Part of Chapter 4 is forthcoming in American Philosophical Quarterly as 'Supererogation and the Case Against the 'Overall Ought'', and an adapted part of Chapter 2 is forthcoming in Analysis as 'Reflective Blindness, Depression and Unpleasant Experiences'.

_____
_____

Signature:		Date:	29/05/2019

Acknowledgements

Figuring out my own views on meta-ethics hasn't been easy for me, let alone for others. It's something that I couldn't have come close to doing without the patience and support from an amazing network of friends and mentors.

Prime place in my acknowledgements should go to my supervisors Alex Gregory and Fiona Woollard. Without their guidance and support I would simply never have been able to produce this thesis. Alex in particular has done a lot to shape my philosophical interests, and (particularly in the first few years) he developed an incredible talent for understanding my own arguments better than I did. Fiona, too, has been an inspiration to me since teaching me ethics when I was an undergraduate all that time ago.

Special thanks also go to Jonathan Way, who played another important role in forming me into the kind of philosopher I am today. Throughout my BA and MA dissertation meetings he would carefully ask "have you thought about <moral philosophy issue> in terms of *reasons?*" A lot of the blame for my thesis on reasons can therefore be pinned on him.

I've also been lucky to be part of an incredible postgraduate community. I owe extensive thanks and gratitude to James McGuiggan, Sophie Keeling, Charlie Boddicker, Eleanor Gwynne, Suki

Finn, Felix Hagenström, and everyone else in the "philosobantz squad" for support, proof-reading, and treasured friendships. I've also had a number of particularly helpful comments from Fionn O'Donovan and Tim Kjeldsen in postgraduate seminars.

I've found the department at The University of Southampton to be overflowing with generosity and advice, not just on my thesis work but on my career and life in general. Among others I should thank Lee Walters for his encouragement with my running (and for distributing helpful philosophy advice as we ran), and Conor McHugh for encouragement and for allowing access to his cat, Aisha, who was a great source of support herself. I would also like to thank Brian McElwee for too many things: including feedback on large portions of thesis work, help in publishing my first paper, and support at conferences. I should also thank him (and Neil Sinclair) for actually reading my whole thesis and then passing me!

As if that weren't enough, a lot of support has also come from outside of philosophy. I'm incredibly lucky to know the amazing Emma Field, Bob Rimington, Lucy Highnett and Bren Markey in particular, all of whom put up with a lot of my nonsense with smiles on their faces. Some thanks even go to my cat Fellini, who has been no support at all.

Finally, I owe extra-extra-special thanks to Will Sharkey. Despite all of his incorrect philosophical views, he has been a source of proof-reading, ideas-discussing, intellectual support, virtue-improving, friendship, and love. He has made me a generous number of chillis and played me at more board games than he would've preferred. Thank you.

Introduction

This thesis will argue that for two concepts with practical *normative force* (namely 'reasons' and 'oughts'), their normative force is contingent on the desires of the agent that the concepts apply to. If the agent doesn't have the necessary desires, then they will not have the corresponding (normative) reasons to act and it won't be the case that (normatively) they ought to act in that way.

When there's something that we ought to do, or something that we have reason to do, there's a certain kind of normative force that exerts itself over us, a kind of authoritative and prescriptive force that weights in favour of an action we might choose to make. I will argue that the source of this normativity is ourselves, our desires. To many of my opponents this will seem like a controversial conclusion, but over the course of my thesis I hope to show that understanding normativity in terms of desire does not mean we have to accept certain radical and implausible claims that have sometimes been attributed to this view. For example, I will show that this view is compatible with there being objective moral principles, with meaningful moral criticism, and with morality as being inescapable. I will show that my view can account for the authoritative feel that normativity has.

My thesis will both give new arguments in favour of this subjective account of normativity and it will provide new defences against objections. I will show the ways that some of the

competing views (such as 'reasons externalism' and a view of 'categorical imperatives') are deficient, in ways that my subjective account is not. And I will show that the mysterious normativity that holds sway over our actions is not mysterious after all, but grounded in the natural phenomenon of desire.

I want to use this thesis to argue that what we have reason to do and what we ought to do are necessarily related to our desires because I want to contribute to the way we understand ourselves, our choices and our moral responsibility. I aim to show during the course of this thesis that such subjectivity isn't a way for agents to escape their responsibilities, but a way to account for them and why they have such a hold on us.

This introduction will begin by explaining more about what I mean by normativity and desire, and what the overall argument of this thesis looks like in a little more detail. After that, I'll give a brief synopsis on the individual arguments of each chapter, together with the context of some of the wider debates that these arguments have their homes in.

What are Desires and Normativity?

The aim of this thesis is, in short, to ground a certain kind of normativity in desires. But normativity is a difficult concept to pin down. I take normativity to be the kind of thing that we find in obligations, in reasons, in actions that we *ought* to take, things we *ought* to believe, the way things *ought* to be. If a concept is a normative one then it comes with a kind of 'force' that tells us that something *ought* to be the case.[1]

Out of the various kinds of normativity that might exist, this thesis is about *practical* normative concepts. That is, it's about actions, and the choices that we make to act in some ways rather than others. This normativity is the kind of force that directs us in some way that we, as agents, can be responsive to, that moves us in some way, or that is the kind of thing that weighs in favour of certain choices that are presented to us. It *pre*scribes what agents should do or should have done rather than *de*scribes. I'll focus on two examples of normativity in particular: *what we have reason to do* in Chapters 1 and 2, and *what we ought to do* in Chapters 3 and 4.

[1] I'll explain normativity in more detail when it comes up later in the thesis. In particular I'll explain normative reasons in **1.1.3** and give more detail about what I mean by the 'normative force' in **3.1.3**.

That, then, is a brief description of what I mean by normativity in the context of this thesis; next, I'll introduce what I mean by 'desire'. This thesis will construe desires very broadly. Perhaps taking them to be something similar to what Williams described as a subject's "motivational set", which he took to be broader than what we might ordinarily refer to as a subject's desires. He said,

> … [a subject's motivational set] can contain such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties and various projects, as they may be abstractly called, embodying commitments of the agent.[2]

Some of these things are examples of what I think should be included in a broad definition of desire, or are at least closely connected. Personal loyalties, projects, and commitments, in particular, seem like the kinds of things that would usually manifest as desires. Dispositions and patterns of evaluation and emotional reaction also seem like plausible candidates.

Perhaps one of the most significant things about the way I want to understand desire is that I include not just desires that are the most vivid and present in a person's mind at any time, but desires that feature in the background of their mind too.[3] When desires cause us to act in certain ways without us really *feeling* them, when we have longer-term desires that aren't prominent in our mind at a given time, and when we aren't even really aware of our desires: these are all the kinds of things I want to include in my broad understanding. This is something that will come up as a theme repeatedly in my thesis.

My talk about desires will be about desires for *states of affairs*, that is, desires for a certain set or sets of circumstances to obtain. For example, it's not just that I desire breakfast, but I have a desire that can be more fully spelled out as wanting a state of affairs in which I have a tasty breakfast in front of me, ready to eat. That doesn't mean desires have to be overly specific, rather I can have a desire for breakfast where there are lots of different states of affairs that I want; there are a large range of types of breakfast that I'd be happy with (some toast, hash brown bap, maybe something with baked beans…), I don't mind whether I eat it at my desk or at a table, and I want it roughly as much now as I want it to appear in ten minutes.[4]

---

[2] Williams, (1981) p.105.
[3] See, for example, Pettit and Smith, (1990) for some discussion of background desires and Schroeder, (2017) who refers to them as occurrent desires (that play "some role in one's psyche" at that particular time) and standing desires (that are not).
[4] This (the desiring for states of affairs) is another phenomenon well described by Schroeder, (2017).

There are several theories about what our desires actually are. Arpaly and Schroeder, for example, argue that to have an intrinsic desire for P "is to constitute P as a reward", and to desire not-P is to constitute it as a punishment.[5] In his Stanford Encyclopedia entry, Schroeder lists action-based theories of desire, pleasure-based theories of desire, good-based theories of desire, attention-based theories of desire, learning-based theories of desire, and holistic theories of desire.[6] According to some theories desires are mental states with a certain direction of fit, ones that aim to fit the world to match the contents of the mind (as opposed to beliefs which are the other way around).[7]

In this thesis I'm largely going to stay neutral on what desires actually are.[8] I take it that it will still be possible to make some meaningful and innovative claims without having *everything* about the concept of desire pinned down.[9] If someone *should* pin the concept down, and an answer should be agreed upon, then even better for me and what I do have to say in this thesis. Until then, I take it to be an advantage for my claims about the relationship between desire and normativity to be compatible with multiple different theories of desire. After all, my arguments will then be plausible to a wider range of people with a wider range of views.

Even if the reader of my thesis does disagree with what I do have to say about desires or my initial position on what it is for something to be normative, then they should hopefully still be able to understand what I mean by these concepts, and, in turn, understand my arguments. I don't think that even in those circumstances my arguments are at risk of being trivial, because it is likely to be the case that many of my opponents will agree with me on my definitions of normativity and desire, but not (yet!) agree with me on the necessary link between them.

Synopsis and Key Debates

[5] Arpaly and Schroeder, (2013) p.127.
[6] Schroeder, (2017).
[7] See, for example, Sobel and Copp, (2001), Gregory, (2012) and Hume (1985).
[8] An exception is in Chapter 2, where I reject pleasure-based theories of desire on the basis that the explanation goes the other way around: desire explains pleasure, instead of pleasure explaining desire.
[9] Not everyone would agree. Aydede, for example, rejects an account of sensory pleasure as being grounded in desire (an account that I'll go on to defend in **2.2**) on the basis that its proponents under-describe what desire actually is. He says it's "…simply not plausible that such an underspecified and open-ended notion can do the heavy-lifting…" in Aydede, (forthcoming a) p.14. But it *can* do a lot of heavy lifting, because there are a lot of things we still do know about desire, and a lot of things we know that it is not. We don't need to know everything to still know plenty.

As I stated above, the overall argument of this thesis is that concepts with *normative force* are contingent on the desires (or sets of desires) of the agent (or sets of agents) that the concepts apply to. I do this in two distinct parts, each focusing on a different normative concept: Chapters 1 and 2 will ground normative reasons for action in desire, Chapters 3 and 4 will ground what we ought to do in desire. Here I'll explain those chapters in a little more detail.

Chapter 1 will begin, like this introduction, with some terminology. There'll be more here than perhaps any other place in this thesis, which is a reflection of the particularly complex webs that writers have woven in trying to understand what it is for something to be a reason for something else. This might be a bad thing to the extent that weaving a path through the literature is pretty difficult, there's a lot of similar terminology that refers to different things, and the concepts sometimes overlap. But overall I think it's more good than bad. I think that there *are* lots of ways to understand what it is for something to be a reason, and starting with a very complicated web of arguments is a necessary step in fully understanding these concepts.

In this chapter I'll go into more detail in explaining *normative* reasons. I'll try to explain what they are in the context of a variety of other kinds of reason, including motivating reasons, explanatory reasons, objective reasons and subjective reasons. In case that wasn't enough reason-terminology, I'll argue that all normative reasons for action are *internal reasons*. This is the term that Williams coined to indicate reasons that are necessarily related to the agent's desires. Again, that should make it clear what role this argument has in the context of a thesis that looks to ground normative concepts in desire.

The positive arguments that I'll give in favour of reasons internalism aren't particularly novel. Rather, I aim to clearly explain the successful arguments of others who've come before me, such as Manne, Markovits, Williams and Goldman. But I'll make some important contributions to the debate not just by making their arguments clear in a larger context, but also by (1) demonstrating that some of Williams' and Goldman's work on reasons internalism is mistaken, and that there is a better understanding of reasons internalism that (2) is immune to an important objection. As the objection goes, the reasons internalist cannot justify why an agent's normative reasons are subjective in some ways but not in others. In correcting the mistakes of a Williams and Goldman –style reasons internalism I'll demonstrate how a reasons internalist can give such a justification after all.

Chapter 2 will continue my defence of reasons internalism, this time primarily against two objections from Parfit. The bulk of the chapter will be about the first argument, which will also be where I make the chapter's most significant contribution to the literature. Parfit argues that what

we have reason to do should not be contingent on what we desire, because what's *valuable* is not based on what we desire, and what's valuable trumps what we desire when it comes to what we have reason to do. He gives the example of pleasure and pain, and argues that we have reasons (other things being equal) to avoid pain and to increase pleasure, but we might not have the corresponding desires. I respond by arguing that pleasure and pain *are* necessarily connected to desire. In fact, what it is for an experience to be a pleasurable one is for the subject to experience a certain kind of desire for it to continue, and what it is for an experience to be unpleasant is for the subject to experience a certain kind of desire for it to stop. This is what Heathwood called the 'motivational account' of pleasurable and painful experiences, and what I will rename the 'desire account'. Chapter 2 will explain one argument in favour of the account, provide a new argument, and respond to a number of counter-examples, including from Bramble and Rachels.

More briefly I'll also discuss and reject what's known as Parfit's 'Agony Argument'. This is a counter-example that's supposed to show that current desires cannot be all that are relevant to an agent's reasons for action, because there might be agents with very unusual sets of desires. When we're faced with reasons either being contingent on an agent's current desires or what's actually better for the agent overall Parfit argues that the latter is more plausible. I'll demonstrate that his example isn't as clear-cut as he thinks, and that if it is then it's a bullet that I'll be happy to bite. In doing so I'll largely be following the arguments of Street.

Chapter 3 moves from reasons to 'oughts'. Here I defend a picture of normative 'oughts' (that is, things that normatively an agent *ought* to do) as 'hypothetical imperatives': imperatives conditional on the agent's desires. I initially defend this idea against 'bootstrapping objections' and 'too-many-reasons' problems, and then argue against one of the account's main rivals: an account of categorical imperatives. I will show here that most of the important features that are characteristically thought of as features of categorical imperatives can be accounted for by hypothetical imperatives too. The only thing that is actually *distinctive* of categorical imperatives when compared to its hypothetical counterparts is that they can apply to agents regardless of their desires. This, I argue, is *not* a feature that a normative 'ought' should have.

Furthermore, I use this chapter to argue that desire-based oughts can actually accommodate many (if not all) of the features of morality that we would want them to. Hypothetical imperatives can have importance and dignity, they can apply to us in virtue of our being rational agents, they can require us to perform actions for their own sake, they can be authoritative and inescapable. Furthermore, a moral system can be made from hypothetical

imperatives and still allow us to have objective moral principles and a meaningful form of moral condemnation.

The second rival to an account of normative oughts as hypothetical imperatives is the target of Chapter 4, and that's what I call a kind of 'overall ought'. Where hypothetical imperatives relate every 'ought' to a desire or set of desires, the kind of 'overall ought' that opposes it might instead be considered to be a kind of intuitive judgment that finds the best overall balance of certain options. In this chapter I give a new argument that such an overall ought is not a plausible concept, because of un-appetising implications about supererogatory acts: acts that go 'above and beyond' what the agent is obligated to do. The best way to understand the 'overall ought', I'll show, is as something more straight-forwardly related to the desires of the agent.

That concludes a brief summary of the chapters to come. The individual introductions for each chapter will contain a more detailed explanation of the work that follows them.

Chapter 1.

What We Have Reason To Do:
A Defence of Reasons Internalism

Introduction

This chapter will argue that normative reasons for action are necessarily contingent on the desires of the agent who has those reasons. This is a similar thesis to what Williams termed 'reasons

internalism',[10] but my own account differs from that of Williams in some of the details. Despite some differences, I consider my account to be a version of reasons internalism. After all, I take the most important part of the theory to be exactly what I just stated: that there is a necessary connection between an agent's reasons and their desires. I'll argue that Williams (and others) are right to say that much, but wrong about in exactly what way they're contingent; in particular, where Williams argued that agents have a normative reason to do what will actually bring about what they desire, I will instead argue that it can at least *sometimes* be about what that agent *believes* will bring about what they desire.

Wider context

My thesis overall argues that for two particular kinds of normative concept, when they are applied to a specific agent, that agent must have a certain, related set of desires. I focus on (1) an agent's normative reasons and (2) what an agent ought to do. The two concepts are related, and I'll explain the nature of this relation in more detail at the beginning of Chapter 3. This means that many of the arguments I give in this thesis will bolster the case for *both* conclusions, supporting the claim that *both* concepts are related to desire. For example, when I argue in **section 3.3.2** that agents only qualify as having moral obligations when they have a certain kind of desire, this will support not only the conclusion that it's directly building towards in the chapter – that what agents ought to do is contingent on their desires – but it will also support the conclusion that what they have reason to do is contingent on desires too. But I should begin somewhere, and this chapter will begin the battle by explaining and defending the reasons-thesis.

This chapter

Despite being the main topic of the chapter, reasons internalism will barely be mentioned until its last **section: 1.3**. The reason for the delay is the extensive set-up that needs to happen first, to make the later arguments clearer. Literature on reasons can be a very tangled web, due to the wide variety of (often overlapping) meanings and concepts that surround the word 'reason'. It wouldn't be so bad if these concepts were all obviously distinct, but they're not. Some of the labels refer to

[10] Williams, (1981). 'Reasons internalism' is the thesis, 'internal reasons' are the reasons (which have, by necessity rather than by coincidence, the connection to the agent's desires.)

concepts that are so similar to others that one might not even notice that they're distinct. Parfit, for example, speaks of the objective and subjective reasons distinction as if it's the same as the internal and external reasons distinction,[11] but I'll show in **section 1.2** that that's not a helpful way to delineate things, because of a second way that reasons can be objective and subjective.

Another reason for the extended set-up in this chapter is that the concepts I'll introduce are necessary for understanding my arguments later. **Section 1.1** will introduce the distinction between explanatory, motivating and normative reasons. It should be obvious why I need to introduce the concept of a normative reason, since I ultimately want to argue that all normative reasons for action *are* internal reasons. Motivating and explanatory reasons are useful concepts in explaining what internal reasons are because they will give me something to compare them with. But it will also be useful to understand what a motivating reason is in particular because some arguments for reasons internalism are about the relationship between those two types: between motivating reasons and normative reasons.

Section 1.2 will explain what objective and subjective reasons are, at least under one understanding of those terms. I'll also argue in this section that there's compelling evidence that normative reasons can't be understood as either fully objective or fully subjective; but rather, we should find ground somewhere between the two, in a way that accommodates our intuitions in the strongest cases. Once again, the explanation of the objective/subjective distinction is valuable for the sake of understanding the literature generally and my place in it, but also as a key part of my defence of internal reasons. In **section 1.3** I'll explain the internal/external reasons distinction more fully and argue that the way that we should draw that line between objective and subjective reasons is not only compatible with an account of reasons internalism but actually equips the reasons internalist with a defence against a prominent objection, attributed to McDowell.[12]

My contribution to the literature in this chapter is primarily two original arguments: the argument in **1.2** about where to draw the line between objective and subjective reasons, and the argument in **1.3** that this way of drawing the line defends reasons internalism against an objection. Much of the rest of the work in this chapter generally will be reviewing and explaining the arguments of others, and setting the scene for arguments in future chapters.

[11] Parfit, (2011a).
[12] McDowell, (1995).

Initial clarifications

Before I begin, I'll make two more initial clarifications about the kinds of reason that are in the scope of this thesis, before I even get around to cutting up the boundaries of reasons within that scope. Firstly, I'll only be concerned with reasons *for action*, or 'practical reasons'. This is as opposed to reasons *for belief*, or 'epistemic reasons', or any other kinds of reasons you might have (for attitudes, for feelings, etc.). Reasons for belief seem to be a different kind of reason to reasons we might have to act. One might think this is because beliefs are less voluntary, or because reasons for belief are more concerned with truth. But that's not something I'll have time to go into in any more detail.

Secondly, I'll restrict myself to talking about 'possessed' reasons. By this I don't mean reasons that are haunted by ghosts, but reasons that are reasons *for* some particular agent or set of agents, rather than reasons that are floating mysteriously unattached (which sounds more like something a ghostly reason would do). Instead of talking about, say, the fact that there generally are reasons to do something (abstract talk of moral reasons, or reasons to push the man in front of the trolley, for example) I want to restrict myself to talk of reasons that 'belong' to specific (even if imaginary) agents, moral reasons for agent A or a reason for agents A, B and C, etc.[13]

Next, then, I'll start untangling the different kinds of reason for action.

1.1 Explanatory, Motivating and Normative Reasons

The first distinction I'll make is between explanatory, motivating and normative reasons. As I mentioned in the introduction, there are two reasons why I want to explain this distinction. Firstly, this is a chapter concerned with normative reasons, and so it needs to be clear what those are and what those aren't. Secondly, understanding these three kinds of reason will also be useful later,

[13] This might be a similar distinction to that between agent-neutral and agent-relative reasons, if there is such a distinction. See, for example, Korsgaard, (1993) and Nagel, (1978). It's worth noting that when I restrict myself in this way I don't mean reasons that are only reasons for *individual* agents. Reasons as I want to pinpoint them could be possessed by any number of agents, although ultimately (as will become clear) this will depend on how many agents have a certain desire or set of desires. I take it that my definition here is broad and will cover most accounts of reasons. After all, for something to count in favour of an action there needs to be an agent (or a set of agents or potential agents) to perform that action.

when I explain arguments for reasons internalism based on the connection between motivating and normative reasons.

1.1.1 Explanatory reasons

Explanatory reasons are simply reasons which explain an agent's action. For example, suppose I go to visit a friend. There might be several different reasons that would feature in an explanation of why I do that; for example I have a reason because I haven't seen them in a while, because they cook great food, and because it'll get me out of the office for a while. It's a very broad term; a lot of different things can feature in an explanation of why someone acts in a certain way. Other examples might be that Bob slammed the door because they were too hungry to think clearly, or that I dropped the glass because I'm clumsy.

Explanatory reasons are the broadest of the three categories. Indeed, they could even describe things which aren't 'actions' as such (but rather habits or non-voluntary movements) or things done by non-agents. For example, Marla leapt out of the way because someone clumsily spilled their drink near her, and the toaster burned the toast because the setting was too high.

1.1.2 Motivating reasons

The category of motivating reasons is more restricted.[14] A motivating reason is a reason why an agent is *motivated* to act in a certain way.[15] But this isn't quite the same as featuring in an *explanation* of why they acted. Explanations, for one thing, don't need to involve the mental states of the agent involved. As Markovits says, "… the fact I haven't gotten enough sleep lately may (partly) explain why I snap at you. But it doesn't *motivate* me to snap at you – it's not the consideration on the basis of which I choose to do so."[16] Motivating reasons are a particular kind of explanatory reason that explains action *in terms* of something about the agent's psychology[17]. So a motivating reason might be that I have a reason to visit a friend because I desire for her to be happy and believe that going

[14] Alvarez also explains and argues for the distinction between explanatory and motivating reasons in Alvarez, (2009) pp.185-186.
[15] Alvarez describes this by saying a motivating reason is one that is "a reason in light of which an agent acts, and it plays the role of a premise in the agent's (implicit or explicit) reasoning…" Alvarez, (2009) p.186.
[16] Markovits, (2014) p.1.
[17] Smith refers to them as being "psychologically real" in Smith, (1987) p.38.

to see her would be a great way to bring that about. Another reason for me to do so might be that I want get out of the office for a while and my friend certainly doesn't live in the office, so visiting her would be a way to do that. These are the things that, to some extent, motivate me to act.

It's worth pointing out here that a motivating reason doesn't need to be for an action the agent actually goes through with. Agents can have motivating reasons that they don't act on, just like how we can be motivated (to an extent) to do things that we never end up doing. I can have several motivating reasons to see my friend, consider them all, and still fail to visit her. Agents can also have mutually incompatible motivating reasons: motivating reasons to do different actions when only one is possible. After all, typically when we're conflicted it's because several options are tempting to us, and we're motivated a little bit to do (or to avoid) each of them. This is a feature that motivating reasons and explanatory reasons actually share: they can explain actions whether actual or possible, and whether past, present or future.

When I listed some motivating reasons above I explained those motivations in terms of something the agent desires. This connection might not be obvious, so now is a good time to mention two things I'll assume in this thesis: (1) that all actions have some motivation behind them and (2) that all motivations have some desire behind them. I'll assume these because I take them to be fairly trivial by the way that I've defined them. For (1), I just take it that part of what it is for some movement to be an action (rather than something done accidentally or out of habit, for example) is for the agent to be motivated to do it. At any rate, not much beyond this section hangs on that distinction. A bit more hangs on (2), but I also take it to be true in virtue of the meaning of the terms as I'm using them. After all, as I explained in the introduction, this thesis takes 'desire' to be a very broad concept, and I include in it whatever it is that does the work in motivating the agent that has it: whatever pushes or pulls them in a certain direction. You might prefer to explain some motivations as being the direct result of, say, perceiving something.[18] But our disagreement is superficial: I simply want to point to a certain feature of how that motivation works and include that in my broad concept of desire.

[18] McDowell (1995) is one source of this kind of thinking. Goldman argues against this position, saying that "… [w]henever externalists employ the perceptual model, holding that we can come to see normative facts as they are, come to see the true values in objects and thereby become motivated to pursue them, the internalist can reply that we could not have become motivated had we not been so disposed, had we not already had related concerns or perhaps hidden character traits." Goldman, (2009) p.14.

1.1.3 Normative reasons

Normative reasons are different yet again. One good way to explain the difference between normative reasons and the other previous two (motivating and explanatory) is this: where motivating and explanatory reasons both describe actions as they are or might be, normative reasons purport to *prescribe* actions. Not to describe things how they are but as how they ought to be, given certain conditions.[19]

Normative reasons for action, then, are reasons why an agent *ought* to act a certain way. Alvarez explains this distinction, saying

> [I]t seems clear that reasons can have normative force. By that I mean that reasons can make something right – not necessarily morally right, but right in some respect. And I do not mean right all things considered but at least pro tanto right. So reasons can be invoked to support claims about what it would be right (for someone) to do, believe, want, feel, etc. (though not necessarily morally right). This feature of reasons underlies a wide variety of the roles that reasons can play, namely, to guide, motivate, evaluate, justify, etc.[20]

Joyce, too, talks about a normative 'force', describing it as a kind of "practical *oomph*."[21]

Some of the reasons described above might also be normative reasons. I have a (normative) reason to visit my friend, for example, such as the reason that it would make her happy, or that it would get me out of the office. And again, like with motivating and explanatory reasons, they don't have to be reasons that I end up following, and the reasons might sometimes conflict. I have reasons to visit my friend, but I might also have several reasons to stay in my office and do some work.

So far I've explained explanatory, motivating and normative reasons. Normative reasons will take a leading role for a lot of this thesis, but a great way to understand them is in contrast to the explanatory and normative kinds. Furthermore, the terminology of the first two will come in handy later.

---

[19] There can also be normative reasons for belief. Although I mentioned above that this thesis restricts itself to talk of practical reasons, there can also be reasons that prescribe what we ought to believe as well as reasons that describe what we do believe or might believe.
[20] Alvarez, (2009) p.182.
[21] Joyce, (2006) p.63.

1.2 Objective and Subjective Reasons

The next distinction I'll cover is that of objective and subjective reasons. As I mentioned briefly in the introduction the labels here are a bit misleading, and I'll begin by explaining why in more detail. I'll then give brief arguments against a particular kind of objectivism and a particular kind of subjectivism, and suggest a way to draw the line between the two views that matches our intuitions in some compelling cases.

1.2.1 Two kinds of objectivity

To put it simply, an objective reason is a reason that's *objective*, that is, external to the agent in some way. Just as there can be objective truths or objective facts (the height of the giraffe might be an objective fact, for example, as opposed to how tall it seems to the other giraffes) so, too, there can be objective reasons. A subjective reason is its counterpart: a reason that's subjective and related to something about the agent in question. Arpaly and Schroeder, for example, describe objective reasons as the kinds of reasons you would have to salt the soup because the soup is lacking salt, and the subjective reasons as the reasons to salt the soup because *you* want the soup to have more salt.[22]

But there are at least two ways that a reason can be objective, and two corresponding ways in which it can be subjective. Suppose that I have a reason to put salt in the soup. If we take this to be a subjective reason, and suppose we further say that I have this reason because I desire to have an adequately seasoned soup. The reason here is dependent on me, the subject, and my desire. If my desires were different, and I preferred my soups to be bland, then I would have no such reason to salt my soup. An objective reason, in contrast, would be a reason that exists *regardless* of what my desires were. For example, one might think that there is a reason to salt the soup (the fact that the soup is bland might constitute this reason) and the reason is just a general reason that might attach itself to the appropriate sets of agents under the right conditions, when it counts in favour of their actions. A more relevant example might be that of objective reasons that apply to specific agents regardless of their desire, such as the moral reasons that we take to apply

[22] Arpaly and Schroeder, (2013) p.53.

universally.[23] But, as you might have already noticed, this is actually the distinction between internal and external reasons that I'm supposed to be putting off until **1.3**. I've introduced it here simply to demonstrate the fact that it's one way (but not the only way) to understand what an 'objective' reason might be, and I'll describe it in more detail in **1.3**.

There's a second way in which reasons can be objective and subjective, and that's in terms of not the agent's desires but their beliefs.[24] Suppose that, for now, we take it that all of the relevant agents *do* want their soup to conform to normal standards of taste. We might think one of these: that an agent has a reason to salt their soup if they *believe* their soup to be unsalted (a kind of subjectivism) or if their soup *actually* is unsalted (a kind of objectivism). This is the kind of objective/subjective divide that I'll refer to when I use the terms.[25]

Next, I'll show that there are good cases to be made against being either fully subjectivists or objectivists about reasons in **1.2.2** and **1.2.3**. Although I think that subjectivism is the more tempting view of the two, I'll offer a way to resolve the debate for those who aren't convinced by either in **1.2.4**.

1.2.2 Against objectivism

Theories about normative reasons face a challenge: understanding to what extent normative reasons should be objective in the sense I described above. When we think about what reasons an agent has to act, we need to be able to understand how much the answer should be influenced by objective facts about what really is the case in the world, and how much the answer should be guided by more subjective facts about what the world seems like to that agent. I aim to show that neither side is without problems (although I'm more sympathetic to the latter), and I'll begin with arguments against understanding normative reasons as objective.

---

[23] I'll go on to reject this kind of reason as being a *normative* reason (and the oughts that follow from it also being normative) in more detail in Chapter 3.

[24] Markovits, too, points out that this kind of objective / subjective distinction is a different one to the internal / external one. Markovits, (2014) p.7

[25] This is also a similar distinction to one called objectivism and *perspectivism* about reasons, although this takes into account things from the *perspective* of an agent, which some might take to be different from what the agent actually believes. For perspectivism, see Littlejohn (forthcoming), Way & Whiting (2017) and Kiesewetter (forthcoming). In a confusingly similar way, Lord refers to the views as objectivism and *perspectivalism* in Lord, (2015).

One of the main arguments against this kind of reasons objectivism is that many of our reasons would be implausibly unrealistic.[26] Agents would have many reasons to do things that they have no idea about, and no way of finding out about, and so would never be able to do. They would have reasons to do things that seem very unlikely to work, things that seem dangerous, and even things that they believe will be harmful.

For example, let's take the following case:

> Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of drugs B and C will completely cure the skin condition; the other though will kill the patient, and there is no way that she can tell which of the two is the perfect cure and which is the killer drug.[27]

There's a fact of the matter about whether drug B or C would cure John. Let's suppose that it's drug C on this occasion. Even though that's the fact of the matter, it's something that Jill has no way of knowing without actually administering the drug, by which point it would be too late to change her mind anyway.

If we understood reasons to be completely objective, then we'd have to say that Jill has the most reason to prescribe drug C. After all, that's the drug that will actually cure John. It has the best projected outcome and will be the best way to bring about the kind of situation that everyone wants: namely, a cured patient who is not dead.

But this result is counter-intuitive. Remember that we're talking here about normative reasons for action. We're not just trying to describe what actions would provide the best outcome, or what action we'd prefer for Jill to take given our objective knowledge. We're trying to work out what Jill's normative reasons for action are, what counts in favour of her acting in a certain way. But it's difficult to see how a doctor can have a normative reason to prescribe a certain medicine if all of the available evidence tells her that it is just as likely to kill her patient as cure him.

[26] In particular I'm considering the reasons that would be different under an objectivist account compared to a subjectivist one. In reality (one would hope) agents will have many of the same reasons under either account, because of having a lot of beliefs that match up to what's actually the case.
[27] Kiesewetter, (2011) and Kiesewetter, (2017).

Considering the risks and the available evidence, any doctor who prescribes medicine C is behaving recklessly, and should not be trusted prescribing drugs at all.

This example is given by others (such as Kiesewetter) and should be enough to establish a problem with this kind of reasons objectivism, but I am also independently suspicious of the theory, for reasons that will become apparent as I go through my arguments in favour of desire-based views of reasons and oughts. For example, normative reasons should, after all, be the kinds of things that agents are capable of acting for (an argument that I'll explain in **1.3.2**), and this is difficult if the agent doesn't *believe* the facts of the matter.[28] So opponents who aren't persuaded by the above case might later be persuaded by other arguments. For now, though, I'll hope that they're at least persuaded enough to be suspicious of complete reasons objectivism.

1.2.3 Against subjectivism

We come across another problem if we take reasons to be completely subjective. This time we can use an example from Williams: it doesn't seem plausible to say an agent has a reason to drink from a mug filled with petrol just because she happens to believe the mug is filled with gin. As Williams puts it,

> The agent believes that this stuff is gin, when it is in fact petrol. [She] wants a gin and tonic. Has [she] reason, or a reason, to mix this stuff with tonic and drink it? (…) On the one hand, it is just very odd to say that [she] has a reason to drink this stuff, and natural to say that [she] has no reason to drink it, although [she] thinks that [she] has. On the other hand, if [she] does drink it, we not only have an explanation of [her] doing so (a reason why [she] did it), but we have such an explanation which is of the reason-for-action form.[29]

If we see someone reaching for a mug of petrol, it seems plausible (as Williams says) to say she has a normative reason to stop, to put it down and get a drink somewhere else.

But the intuitions here don't match up with what we'd have to say about the agent's reasons if we were to be completely subjective about reasons. The agent believes that the mug is filled with

[28] See also Manne, (2013).
[29] Williams, (1981) p.102.

gin, and presumably wants some gin to drink after a long, hard day of writing her thesis. For the subjectivist account that's enough for us to say that she has reason to drink the stuff in question.

As I've said above, I'm more persuaded by subjectivism than I am by objectivism, for reasons that are independent of these two examples, but to do with arguments that might run similarly to arguments in favour of reasons internalism. Arguments, for example, to do with the importance of connecting an agent's normative reasons for action with the agent's moral psychology, with reasons that the agent is capable of acting for. The arguments in **1.3.2** which talk about an agent's capacity for action might apply to an agent's beliefs as well as their desires.

But there might be disanalogies between the arguments for reasons internalism and reasons subjectivism, and in case there are, I want to concentrate on defending the former. To make my case for it stronger I'll aim to defend reasons internalism while staying as neutral as I can on the matter of reasons subjectivism. If that theory is true then all the better for me, but if you find examples like the gin case to be persuasive then, in **1.3.4**, I'll show that reasons internalism is by no means ruled out.

1.2.4 Where to draw the line

If, then, my opponent *does* find these counter-examples to be compelling, I'll next suggest a good way to draw the line between objectivism and subjectivism, to help determine what's relevant when considering why our normative reasons should sometimes take into consideration the beliefs of an agent rather than the facts of the matter (or the other way around). It might be that a good account of normative reasons is one that accounts for the differences between cases like those above: those where objective facts do seem to influence normative reasons and those cases when they don't. This might be particularly the case for examples like the two that I listed, where our intuitions are fairly strong (as philosophical intuitions go).

My proposal is this: a fact influences what normative reasons an agent has if and only if that fact is something that it would be easy to persuade the agent of.[30] I'll go on to show that this has several advantages, and gives us intuitive answers to both the doctor and the gin cases listed above.

[30] This 'persuasion' approach is not completely novel; appealing to an 'ideal advisor' has been used as a tool for understanding reasons by others such as Smith, (1994), Manne (2014) who uses it to argue in favour of reasons internalism, and Bennett, (1997) who talks about appealing to a well-informed bystander.

Suppose we return to 'the gin case': the agent reaching for the mug filled with petrol because she believes it to be a mug of gin. In most of the ways we might fill out the details of this case it would be relatively easy to persuade the agent of the objective fact: that the mug doesn't contain what she thought it did, but rather petrol. Because it's something so easy to persuade her of, then according to the persuadability account then it's a clear case of something that does influence what she has normative reason to do. Her reasons are influenced by the facts of the matter because it would be easy to persuade her of them.

Next we can return to 'the doctor case': the doctor who has to prescribe medicine. There is no evidence available to her about which drug of B and C will kill and which will cure her patient. Intuitively the doctor does not have a normative reason to prescribe the drug that would objectively do the best. In fact this would be positively reckless. The persuadability account agrees with this verdict: an imaginary advisor would find it difficult to persuade the doctor of the objective fact, so it doesn't influence her normative reasons. The imaginary persuader doesn't, after all, have any special rhetorical skills or the ability to present her with any new evidence. The facts of the matter, then, wouldn't be able to trump the doctor's beliefs in the matter, and her beliefs will be what her normative reasons are contingent on.

An account like this one can successfully differentiate between the two kinds of cases, and next I'll go further in explaining how. The imaginary persuader is successful in hypothetically persuading the agent in the gin case but not the doctor case, but this difference isn't arbitrary, and it's not the case that the imaginary persuader is working any differently in the two cases. Rather, it's a way to show (among other things) that there's evidence that's more available to the agent in the gin case. The doctor has no way to determine which medicine would be the successful one, there are no tests that she could run, no medical journals she could peruse for the answer. The thirsty agent has no idea that her mug is filled with petrol, but there are some very simple steps she could take to find that out. She could smell or look a little more carefully. She could consider what she knows about how the mug got there and came to be filled with a liquid. The imaginary persuader isn't providing the agent with a completely new source of information, but (in certain circumstances) bringing to light what's already there and already easy for the agent to access herself.

Persuadability is a good basis for determining where to draw the line between objectivism and subjectivism about reasons because it tracks a variety of things that seem to influence our intuitions in these cases: whether there is easy evidence that we can point to, what the probabilities are, and how risky the situation is. For example, the gin case has easily accessible evidence and the doctor case does not. In the gin case the bigger risk would be if she drank the petrol, in the doctor

case the biggest risk would be prescribing one of medicines B or C and being incorrect. These would affect both our intuitions on normative reasons and of persuadability.

Another reason why this account provides us with a good answer is this: it tracks our intuitions about reasons because we often think about normative reasons in terms of advice and persuasion anyway. The gin case seems counter-intuitive because we want to be able to advise the agent of her reason to act differently, to *persuade* her not to drink from the mug. We want to be able to point out to her that she has a reason that she didn't realise she had. We wouldn't be able to do that in the doctor case, certainly not without a lot of things being different about the world. These are the kinds of things which seem to influence our intuitions about normative reasons, and persuadability can be a way that our account of normative reasons matches onto them as best as it can.

I'll address one more point about the persuadability account before I move on to reasons internalism. At first this might sound like a concession, rather than a clarification: the account is vague. It tells us how to track which features influence normative reasons, but it does so in a way that doesn't give us any clear answers in a lot of cases. The doctor case and the gin case are atypical; they're the two cases that provide us with some of the strongest evidence against each of the two accounts. But generally there's no determinate answer to which things will be 'easy' to persuade an agent of and which will be difficult to. The account tells us something about normative reasons without by any means telling us the whole story. But, as frustrating as it might be not to have an exact formula, this vagueness could also be seen as a virtue, because our intuitions about normative reasons are also vague. We cannot know in advance which cases will give us which intuitions, and there is going to be a lot of disagreement. But we don't need persuadability to be a tool to work out cases that we're uncertain about, we just need it to provide a promising way of understanding the most obvious cases. Something that would lead us down the right path, that points us in the right direction.

Talking about our intuition on reasons, Joyce says: "if a philosopher sets out to analyze or explicate a concept in ordinary parlance – like having a reason – then she must start with how the word is generally used."[31] This is something that the persuadability account does, by giving us as close as we can get to a specific account while still paying respects to how we use and talk about normative reasons. And if we then did want to work out the specifics, this account gives us a good place to start.

[31] Joyce, (2001) p.102.

In this section I've argued, briefly, that an account based on persuadability is one good way to satisfy difficult intuitions about a certain kind of objectivity and subjectivity about normative reasons for action. It's important to note here that my argument on persuadability isn't an attempt to argue that persuadability is a necessary feature of the concept of a normative reason. I don't think that it's integral to our understanding of the concept to picture some kind of omniscient persuader, for example, and nor do I want to claim that it's this imaginary persuader who is really doing the work in drawing the line between the different kinds of reason. What I do mean to have done is to suggest a way of tracking the most relevant features of our normative reasons intuitions or language which will be influencing us in these extreme cases. For the most obvious examples the persuadability seems to track the right answers, and although it's not as helpful in the grey areas I don't think it *needs* to be able to provide us with the exact answers.

For the purposes of my project, it's not as important to understand whether beliefs or facts influence what an agent's normative reasons for action are. But what I do aim to have done is to have argued that for anyone persuaded by the 'common-sense' intuitions in the two extreme cases, then there is a possible solution. More importantly, in **1.3.4** I will go on to show that this way of understanding these problem cases will solve what would have otherwise been a tricky difficulty for reasons internalists.

1.3 Internal and External Reasons

Finally, I'll turn to the distinction between internal and external reasons. I'll explain what reasons internalism is in **1.3.1**, and briefly explain some of the best arguments in its favour in **1.3.2**. In **1.3.3** and **1.3.4** respectively I'll introduce a potential objection and then a solution to it based on the work done in the previous sections.

1.3.1 Reasons internalism

Reasons internalism is the thesis that an agent's normative reasons to act are necessarily connected to what she currently desires. The exact nature of that connection is controversial, as should be clear from the discussion of the objectivism / subjectivism divide in the previous section. It might sometimes be the case that an agent's normative reasons will be reasons to act in ways that *will*

bring about states of affairs that the agent desires, or it might be just that they are the reasons to act in ways the agent *believes* will bring about the states of affairs that the agent desires. But for now I'll put the objectivism / subjectivism divide aside, and concentrate on the contingency on desires.

Williams introduced the internal and external reasons debate, and these two different ways of understanding what a reason is. He says,

> Sentences of the forms 'A has a reason to φ' or 'There is a reason for A to φ' (where φ stands in for some verb of action) seem on the face of it to have two different sorts of interpretation. On the first, the truth of the sentence implies, very roughly, that A has some motive that will be furthered by [their] φ-ing, and if this turns out not to be so the sentence is false: there is a condition relating to the agent's aims, and if this is not satisfied it is not true to say, on this interpretation, that [they have] a reason to φ. On the second interpretation, there is no such condition and the reason-sentence will not be falsified by the absence of an appropriate motive. I shall call the first the 'internal', the second the 'external' interpretation.[32]

An internal reason, then, is one that is connected in this necessary way to the desires of the agent, and an external reason (if it exists) is one that has no such necessary connection. It's not the case that external reasons are only those which *don't* have a connection to an agent's desires, but rather any reasons where the connection is not necessary for the reason to still be a normative reason for that agent to act.[33] A certain external reason might happen to match the desires of an agent, for example, but it wouldn't need to. For the purpose of discussing the differences between the two theories, though, most of the external reasons I'll discuss will be those which aren't connected with the agent's desires.

Before moving on to arguments for reasons internalism I will make a few more clarifications about internal reasons in practice, using examples. Firstly, internal reasons needn't only be reasons to achieve what one is consciously thinking about desiring, or to satisfy particularly obvious or strongly-held desires. An agent could also have internal reasons to fulfil desires she hadn't even considered at the time, as long as she still held them. Suppose we think about an agent named Kima who's angry with her brother, and is so overcome with anger at something he's done that she feels nothing at a particular moment other than a strong desire to hit him. If we were to list the (internal) reasons she had, then it wouldn't be the case that she'd only have reasons to hit

[32] Williams, (1981) p.101.
[33] Parfit makes this point in Parfit, (1997) p.104.

him, because the desires most at the forefront of her consciousness aren't the only ones she has. There are likely to still plenty of other desires in the background of her mind that will give her reasons to restrain herself, and these desires are not necessarily less important for being *felt* less strongly at that particular moment. Kima may also have desires to stay out of trouble or to be a good role model, for example, which are just not at the forefront of her thinking after her brother has angered her. Behind the anger she will even still care about the welfare of her brother. All of these things that she desires will give her internal reasons to refrain from hitting him.

Another example here is one I'll borrow from Foot. She says,

> Sometimes what a man should do depends on his passing inclinations, […] Sometimes it depends on some long-term project, when the feelings and inclinations of the moment are irrelevant. If one wants to be a respectable philosopher one should get up in the mornings and do some work, though just at that moment when one should do it the thought of being a respectable philosopher leaves one cold.[34]

This is a plausible description of what our desires seem to be like. When the only thing that our explicit thoughts are focused on is a certain desire, that doesn't mean it's the only desire or even our strongest desire at a given time. We wouldn't say of the distracted or sleepy agent that they don't want to be a philosopher when they're feeling like that. In fact, what makes situations like these really difficult is the fact that vivid desires are competing with other longer-term but less strongly felt desires. Whenever we resist strong temptations like these we do so because of those other desires, like the desire to be a good philosopher or a good sibling.

I'll make a second clarification here. Internal reasons can be moral reasons, too. Suppose Kima desires to be a good person. This gives her a reason to do good things.[35] It gives her a reason to keep certain promises, to feed the homeless and to put the welfare of others before herself. This is the case for nearly everyone, since (fortunately) nearly everyone has a desire to be good.[36]

Let's contrast internal reasons with external ones again, for a clearer picture. Suppose Kima's girlfriend, Cheryl, has no such desire at all. Cheryl does not have an internal reason to act

---

[34] Foot, (1972) p.306.

[35] There might be issues here with what kinds of desires might actually be the desires that a good person has. For example, it might be the case that good people aren't good in virtue of a desire to be good, but good in virtue of a desire to help others, a desire to bring about good consequences, a desire to obey the moral law, etc. In fact, some people might argue that simply desiring what's good because of the fact that it's good is "fetishistic". See, for example, Smith, (1994).

Any of these desires would fit with the kind of reasons internalist view I discuss here.

[36] I'll have more to say on this point in **3.3**, but this point is also made by Brink, (1989).

well, since she does not have the requisite desire to be good, even in the background of her mind, or as a long-term preference. Someone who believes that (at least some) normative reasons are external reasons may say that Cheryl and Kima nevertheless both have the same reasons to be good people and perform good actions, because these reasons exist and apply to them no matter what desires they each have.

The internal reasons theorist has to put their foot down and say that Cheryl doesn't have the same reasons to be good as Kima does. But they still have a lot of room to work with in terms of the kinds of reasons we *can* describe agents like Cheryl as having. It might be useful to briefly examine a wider variety of internal reasons to make sure the picture is clear. For example, even though Cheryl has no intrinsic desire to be a good moral person, she may still have other relevant desires that we can use to persuade her that she does have a reason to do good things. Suppose an opportunity arises for both Kima and Cheryl to donate some money to charity. Kima sees that she has a reason to donate some money because she knows that doing so would be the good thing to do, and doing what's good is something that she wants. For Cheryl, the opportunity to become a better person (or to do what's good) doesn't directly motivate her to donate the money but other considerations might, ones that relate closely to it. For example, donating money to charity and appearing to be a good person will make Kima like her more, and Cheryl definitely has a desire for Kima to like her. It might also set a good example for others around her to be more generous in future, and although Cheryl doesn't see the direct benefit of being generous herself she may desire to bring about that behaviour in other people. After all, if she is ever in need of charity she might be better off in a community of people who have learnt to be generous by seeing such generosity in others.

Another way to work with agents like Cheryl – and a way that's compatible with a reasons internalist picture of things – might be to try to cultivate the right kinds of desires in her, and to encourage and promote activities that are likely to bring about good desires in her. This seems like a plausible way to bring about good people, insofar as that might be possible. Arpaly and Schroeder, for example, describe virtuous agents as those with the right desires.[37] This lines up nicely with a reasons-internalist picture of things; after all, it makes sense that the things one might be able to do to make someone have good moral desires would coincide with the right kinds of things one might be able to do to give someone reasons to act in ways that are good.

What I've hoped to demonstrate using the above examples is that internal reasons, despite being necessarily tied to an individual agent's desires, can be incredibly varied. That is, they can

[37] Arpaly and Schroeder, (2013). This is another theme that will return in more detail in Chapter 3.

include reasons that relate to an agent's less immediate and obviously felt desires, they can include moral reasons, and they can include reasons that the agent herself may not be aware of, which an interlocutor might be able to persuade her of after some discussion. They can do all of these things without us needing to appeal to external reasons.

1.3.2 Arguments in favour

Next, I'll briefly explain two arguments in favour of reasons internalism. Although these are largely arguments made elsewhere and by other people, my explication here will be useful both in giving the reader an idea of the motivations behind a reasons internalist picture, and build a brief foundation for the arguments in future chapters to build on. Furthermore, the arguments in favour of reasons internalism, as I describe them, leave the theory apparently vulnerable to a particular objection, which is what I'll tackle in the remaining sections of this chapter.[38]

Arguments about capacity for action

The first argument for reasons internalism that I'll discuss comes in several versions, which I'll run through now. They are versions of the argument that a normative reason for action, when it's a reason *for* a specific agent or a set of agents, needs to be able to appeal to the agent's psychology in a way that means those reasons might, without the agent going through too many changes, move that agent to action. As Goldman puts it, "Reasons must be capable of motivating us."[39] This argument in particular draws on the distinctions that I made in **section 1.1**, between explanatory, motivating and normative reasons. Goldman, again, says the following,

> …[R]easons must be explanatory and not simply normative if we ever do act on them. And they cannot be normative unless they are also potentially explanatory, since there is no point in telling people they ought to act on certain reasons unless they can act on them, unless the reasons are able to motivate them.[40]

[38] My method will be to run through the two arguments I find most convincing, but for other summaries of the different kinds of argument in its favour see Brunero, (forthcoming) and Heathwood, (2011).
[39] Goldman, (2009) p.8. Goldman in particular focuses on the connection between reasons and motivations to support reasons internalism, but it's important to emphasise that reasons internalism doesn't require that all reasons must actually motivate, even if it's only to a small extent. It's just that they must in some sense be *capable* of motivating the agent. I'll explore this distinction in more detail in the rest of this section.
[40] Goldman, (2009) pp.29-30.

Another way of arguing a similar point is that it seems that part of what it is to be a normative reason for an agent to act is that it needs to be possible, given some use of the word 'possible' at least, for that reason to cause the agent to act *for that reason*. Williams says something similar,[41]

> One reason why [a reason must be able to explain an action] is that it plays an important part in discussions about what people should be disposed to do. […] Taking other people's perspective on a situation, we hope to be able to point out that they have a reason to do things they did not think they had a reason to do, or, perhaps, less reason to do certain things than they thought they had.[42]

Suppose I have a reason to attend a protest because I want to stand against refugee detention centres. It seems plausible that this is a normative reason for me because it's the kind of reason I could act for. I might not choose to attend the protest, but if I did then the fact that I want to stand against refugee detention centres is the kind of thing that would've featured in my motivation; it's the kind of reason that I could have acted *for*. This might be the case even if the reason wouldn't be sufficient to move me to action on its own. I might, for example, only ever be moved to attend the protest if there were several reasons in favour of me going.

Williams compares this to how an external reason would work. He says,

> The whole point of external reason statements it that they can be true independently of the agent's motivations. But nothing can explain an agent's (intentional) actions except something that motivates him so to act.[43]

A reason unconnected to the desires of the agent, then, could not be the reason for which the agent acted. Suppose we said that I have an external reason to attend the protest because it's being held near a particular pizza chain. If I don't, on any level, have desires that relate to being near the pizza chain (if I don't, for example, like the pizza), then it's not the kind of reason that would be the reason for which I act. Even if I do end up going to the protest, it won't be for *that* reason.

---

[41] Finlay, (2009) describes Williams' argument as being that it's part of the *concept* of a normative reason that it should be able to explain, in some sense, the agent's action.
[42] Williams, (1981).
[43] Williams, (1981) p.107.

Markovits describes Williams as making another, similar, point. She says,

> The whole point of ascribing a reason to someone, either internal or external, Williams thinks, is to make clear to them that if they fail to act accordingly, they are failing by their own lights – they are failing to live up to a standard whose bindingness on them they must themselves, as rational agents, acknowledge: the standard of rationality. [44]

So another reason why we might want to think that normative reasons must be capable in some sense of motivating agents is that normative reasons should refer to the standards *of the agent in question*. The normative force isn't something that can only be externally imposed, but must be in some sense internal to the agent, something that they can come to see (or how else would they ever act for that reason?)

Markovits makes more points in favour of this kind of argument. She says that internal reasons are important because they help prevent agents from being "alienated" from their reasons.[45] Furthermore, the connection between reasons and desires allows that your reasons actually be action-guiding.[46]

Arguments about external reasons as a different phenomenon

The second argument, and one that I find to be one of the most compelling, is the following. When we talk about reasons that *aren't* related to our desires it seems like either one of two things is the case: either we're talking about something different to what we do when we're talking about normative reasons, or we're trying to talk about normative reasons but failing.

To make this argument as clear as I can, I'll start by going through the uses of normative reasons. When we talk about normative reasons we mean something that, as I mentioned above, *could* possibly motivate someone to action. There's some kind of 'normative force' that goes with the reason, something which means it 'counts in favour' of the action for the agent in question. We use these reasons to deliberate, to reflect, and to persuade and reason with others. Manne puts it like this:

---

[44] Markovits, (2014) p.35.
[45] Markovits, (2014) p.54. Her talk of alienation is meant to be similar to the kind of alienation Railton talks about in Railton, (1984). Manne also picks up on this connection in Manne, (2014) p.97.
[46] Markovits, (2014) p.54.

> Think about our practices of talking to each other, and reasoning with each other, as well as by ourselves. Think about more than that, too, though: think about the ways we instruct, reproach, request, cajole, wheedle, manipulate, demand, condemn, yell, and even stamp our feet on the ground in disgust at people's conduct. Think, in other words, about the whole teeming mess of embodied and socially-situated normative behavior—i.e., behavior by means of which we give voice to ideas about what to do, and also what should happen.[47]

All of these ways that we use normative reasons could only succeed if they appeal to the actual agent's wants, cares, preferences. Her *desires,* broadly construed. This is also a useful point to remember for the first argument. As Goldman says, "…there is no point in telling people they ought to act on certain reasons unless they can act on them, unless the reasons are able to motivate them."[48]

This is the case even if we don't take "moving the agent to act" as the success-condition for the above uses of normative-reasons. There are still genuine cases of reasoning with people that might not actually persuade them to act in a certain way. Some of it, for example, might be retrospective reasoning. Even in these cases it seems like we're only actually successfully speaking about *their reasons* if the reasons that we're talking about have a connection with the agent's actual desires at the time.

This argument against external reasons is made by a number of reasons internalists, including Williams. He says:

> If an agent really is uninterested in pursuing what he needs; and this is not the product of false belief; and he could not reach any such motives from motives he has by the kind of deliberative processes we have discussed; then I think we do have to say that in the internal sense he indeed has no reason to pursue these things. In saying this, however, we have to bear in mind how strong these assumptions are, and how seldom we are likely to think that we know them to be true. When we say that a person has reason to take medicine which he needs, although he consistently and persuasively denies any interest in preserving his health, we may well still be speaking in the internal sense, with the thought that really at some level he *must* want to be well.[49]

[47] Manne, (2014) p.94.
[48] Goldman, (2009) p.29-30.
[49] Williams, (1981) p.106.

When we try to reason with someone without appealing to their desires, Williams says, it seems right to say that we're making an optimistic mistake about the kinds of desire that agent actually has. He then says:

> The sort of considerations offered here strongly suggest to me that external reason statements, when definitely isolated as such, are false, or incoherent, or really something else misleadingly expressed. […] Those who use these words often seem, rather, to be entertaining an optimistic internal reason claim, but sometimes the statement is indeed offered as standing definitely outside the agent's [motivational set] and what he might derive from it in what is meant. Sometimes it is little more than that things would be better if the agent so acted. But the formulation in terms of reasons does have an effect, particularly in its suggestion that the agent is being irrational, and this suggestion, once the basis of an internal reason claim has been clearly laid aside, is bluff. If this is so, the only real claims about reasons for action will be internal claims.[50]

Manne also lists more things we might be doing instead of offering a genuinely normative reason in these situations,

> I suggest that it is only when we relate to other people as such, thus adopting the interpersonal stance towards them, that we can be said to reason with them. This is as opposed to ordering them about, coercing them, or trying to 'manage' their behavior (among myriad other possibilities).[51]

And, later, she says, "[w]hen we are trying to convince someone to do something and they have no relevant desires, what we're trying to, when we were previously trying to reason, seems to change."[52] Finally, she makes this point (in a long quotation, but the point is important and well-put):

> I certainly do not believe that we have to retreat from doing or saying anything, if we discover that the callous husband has no motivational propensity to treat his wife more nicely. Rather, I believe that, if we discover that there is no way of motivating him to do so, simply by reasoning with him, then we are

[50] Williams, (1981) p.111.
[51] Manne, (2014) p.91.
[52] Manne, (2014) p.103.

> no longer well-described as trying to offer him a reason. He is beyond the reach of such reasons, at least as things currently stand. We are consigned to doing something more in the vein of giving him an order, or simply expressing our disapproval. It can be important for us to state 'for the record' that we find both his actions and his attitude unacceptable. We may also have to resort to trying to manage this man's behavior—by sending him to anger management class, or helping his wife to get out of there, or assisting her in obtaining a restraining order against him. Or we may have to try to get him arrested, hoping that he will be locked up, at least until he can be reformed and hopefully rehabilitated. Any of these interventions might well be the sort of thing which we ourselves have good or decisive reasons to do. Our hands are not tied. But this man's motivational profile should make a fundamental difference to our sense of where we stand in our moral-cum-social relationship to him. Our stance towards him is (or, at least, should be) no longer interpersonal, at least to the extent that we are still trying to influence the way he treats his wife. Rather, it becomes (or, again, should become) objective. In this sense, our tongues are tied by his motivational deficits. This is admittedly sad, but it may be true nonetheless.[53]

Out of all the things that we do when we use 'external' reasons, ones that don't make reference to the desires of the agent in question, none of those things seem to plausibly be a case of using *normative reasons*. After all, we're talking about normative reasons for an agent to act, that are reasons that count in favour of an action *for them*. And how could a reason count in favour of something for someone if that someone in no way favours that thing?

In this section I have given two arguments why one might find reasons internalism appealing. The first was all related to the idea that an agent's reasons, even her normative ones, must be in some way related to her psychology, in that they must in some sense be actions she is capable of performing.

The second argument was made by demonstrating that when we appeal to reasons *without* the connection to their desires being there, without the actions being something that the agent could possibly do, we seem to either be mistaken or be talking about something different. I listed a number of things that these 'external reasons' statements could be aiming to do, borrowed from a variety of authors. This included trying to use rhetoric to give the agents new reasons to act, and attempting to order, coerce or manage them.

1.3.3 A problem

[53] Manne, (2014) p.110. An example of someone making similar suggestions is Finlay, (2014) pp.183-185 in particular.

Reasons internalism is, as I've shown, the account of normative reasons in which an agent's reasons are those which relate to her desires, broadly construed. As I've shown above with my brief discussion of the first argument in favour of reasons internalism, the connection, roughly, ensures that there is some sense in which an agent is *capable* of acting for her normative reasons, that they really are something that will weigh in favour of an action *for her*.[54]

This kind of argument leaves the reasons internalist vulnerable to a certain objection, and it's that I'll turn to next. The objection is this: there are persuasive reasons to think that an agent's normative reasons aren't entirely based on their beliefs, but rather based to some extent on what's actually the case in the world. If we need to 'idealise' an agent's actual set of beliefs in this way to determine what her reasons are, to determine why they're reasons *for her*, then the reasons internalist needs to justify why we shouldn't do something similar for the agent's desires. Instead of an agent's reasons for action being contingent on what she actually desires, why aren't they contingent on what she *should* desire, or *might* desire under some idealised conditions?[55]

This would be a problem for reasons internalism because according to reasons internalism an agent's reasons are contingent not on some ideal or hypothetical set of desires,[56] but on the agent's actual, current set of desires. Appealing to anything other than the agent's actual desires would undo the point of reasons internalism as being something that appeals to something specific about the agent's actual psychology.

It's worth noting that both Williams and Goldman's versions of reasons internalism are vulnerable to this kind of objection. Williams, for example, gave the original gin example in his internal and external reasons paper, and argued that we should take into account the truth of the matter rather than the agent's beliefs.[57] Goldman argued for the same thing in his own version of reasons internalism, on the grounds that acting in accordance with your desires might otherwise be self-defeating.[58] To idealise the agent's beliefs in this way (so that only the facts or the true beliefs count) without idealising the desires seems like a move that the reasons internalist needs to justify.

---

[54] By 'acting for a reason' here I mean the kind of thing that Davidson meant when he spoke of reasons 'rationalising' actions in Davidson, (1963).
[55] This objection can be attributed to McDowell in McDowell, (1995). It's also been discussed as a problem very clearly by Mason, (2006).
[56] For an alternative view of reasons that does try to appeal to an idealised set of desires see Smith, (1994).
[57] Williams, (1981).
[58] Goldman, (2009).

1.3.4 A solution

In **1.2** I discussed whether or not an agent's beliefs should be taken as they are when determining an agent's reasons, or whether they should be 'idealised' in some sense, and the facts of the matter should be taken into account instead. I explained that I have sympathies for the former view: that normative reasons are subjective when it comes to the agent's beliefs. But I went on to argue that if there should be some middle-ground between subjectivism and objectivism about beliefs then a good way to locate this middle-ground would be by considering whether the facts of the matter (when they're different from what the agent believes) are things that, hypothetically, the agent could be easily persuaded of. This kind of approach seemed to track a selection of variables that seem relevant to what an agent's reasons are, such as the risks involved and the availability of evidence.

In this section I'm going to explore the persuadability approach in terms of an agent's desires instead of their beliefs. I'll argue that if persuadability is a good way to determine what facts about an agent need to be 'idealised', then this provides the reasons internalist with a good justification for the asymmetry between desires and beliefs. New desires, I will argue, are not the kinds of things that an agent can be easily persuaded of.

In the clear-cut cases, I argued, an agent could be persuaded of the fact of the matter when the risks of not adjusting their belief would be high, for example, or the evidence was such that she might easily come to recognise or see it herself. When it comes to desire, one of two things will be the case when an imaginary advisor could come in to help us to determine what their reasons are. Either an imaginary persuader would be easily able to persuade the agent of a desire because they already have it or they wouldn't, easily, be able to persuade the agent of that different desire at all.

Reasons internalism as I've understood it, after all, takes desires to be a very broad category. As I'm not shy about emphasising, they include standing and latent desires as well as those more vivid and obvious to the agent at the time. It seems like any case analogous to the gin case would involve the imaginary persuader bringing to light desires that the agent already has, but perhaps feels less strongly at the time. Suppose that Fellini can only concentrate on his desire to stay in bed. An imaginary persuader might easily be able to remind Fellini of his love of philosophy and related desires in order to bring those to his mind. Following the persuadability account, then, and if there were a symmetry between desires and beliefs, then Fellini would have a reason to get out of bed in order to do his work. But this is compatible with reasons internalism, because he still

has the relevant desires, it's just that at the time they're in the background of his thought, not the foreground.

Let's consider a case in which the subject doesn't have any of the relevant desires at all. Fellini, for example, does not want to join the army. He doesn't believe in the military, he doesn't want to fight for his country, he doesn't think that it will improve him at all as a person and he has no familial pressure. This example is far closer to that of the doctor case because there's no easy way to give Fellini the relevant desires.[59]

It seems that this would be the case for all desires. Where in some cases there might be easy ways to bring about new beliefs in an agent, there doesn't seem to be an equally simple way to bring about new desires in an agent.

I'll discuss an objection next, based on whether it would it be easier to persuade an agent of insignificant desires. In the belief case it would be easier to persuade an agent of a new belief if there's no real risk that comes from adopting that belief. The doctor would've been very difficult to persuade, for example, because the risks were high: a patient would die if she made the wrong decision. There might be cases in which the agent's beliefs are far less risky and the persuasion is easier. Perhaps this kind of case would be an easier way to find an analogy with desires.

Suppose that Cressida, for example, doesn't want to try the coconut sorbet that Lola has just offered her. It's not that she hates either coconut or sorbet, just that she feels nothing about the flavours. She has no feelings either way about trying new things, or about new flavours of sorbet. Here it might well be fairly easy to persuade her to adopt these desires. Adopting them would certainly involve minimal risk. And, just as we saw in the gin case, it might be possible by simply bringing to light something that's already available to her. Perhaps we could just imagine the imaginary persuader telling Cressida that she'd really enjoy the sorbet, and that would be enough.

My response here is to point out that despite first appearances this isn't a case of giving Cressida new desires. The imaginary persuader here isn't presenting her with new desires, but new beliefs. She already broadly desires experiences that she'd really enjoy, for example, so the imaginary persuader isn't giving her a new desire as much as informing her of a new way to fulfil the desires she already has. So this isn't a problem for reasons internalism.

[59] This case is similar to one that Williams discusses in Williams, (1981).

This kind of result matches one of the things that seems to be important about reasons internalism, according to the first argument I described in **1.3.2**: the relationship between our reasons and our capacity for action, for motivation. Our easily-changed beliefs aren't important to our personality, our psychology, etc., but our desires are. This is both why it's difficult to persuade agents of new desires, and why different desires shouldn't factor into what an agent's normative reasons are.

Other than by presenting the agent with new access to their current desires (or the ways to attempt to satisfy them), there doesn't seem to be a way to bring about new desires in an agent. But that's not to say that agents never do get new desires. There are lots of ways in which agents can come to appreciate and desire things that they'd never desired before. There can be transformative experiences, for example. Agents can have experiences that change their view on the world, or that give them new values. They can become bored of certain things and excited with others. But this is all compatible with reasons internalism. After all, gaining those new desires will give agents new reasons to act. But this section aimed to establish that gaining such new desires isn't *easy*, and certainly not as easy as it is in examples like the gin case. If the reasons internalist wants to concede that not all normative reasons are subjective in terms of belief, then, they still have a good answer for why normative reasons are subjective in terms of desire.

In this section I defended reasons internalism against an objection. In doing so, I gave the reasons internalists a way to justify how their concepts of a normative reason can fit in with how we linguistically use the term. Reasons internalists can, if they want to, claim that normative reasons are not always contingent on a subject's beliefs, but rather on the facts of the matter in certain cases. And they can do so while justifying the asymmetries with the case of desires. They can appeal to the account of persuadability, and say that agents have a reason to act if they could be easily persuaded to act in that way. Reasons internalism, I've argued, would follow from this, and it provides us with intuitive answers in problematic cases about an agent's beliefs.

Conclusion

This chapter had several aims and several arguments, which I'll briefly recap here. Firstly, my goal was to untangle some of the different terminology about reasons. I clarified that this thesis is interested in reasons that are practical (they're reasons for an agent to act, rather than for an agent to believe, for example) and they're reasons for a specific agent. In **1.1** I then explained the

distinction between explanatory, motivating and normative reasons, and said that my interests lie with the latter kind: the kind that are about what an agent *should* do (or have done), rather than reasons that *only* aim to explain the actions or motivations of the agent.

In **1.2** I began discussing one way in which reasons can be objective or subjective: in terms of beliefs (that are subjective) or facts of the matter (that are objective). I argued that reasons should not be completely objective in this sense, because of cases like that of the doctor who needs to prescribe medicine but doesn't have enough data to know which medicine is objectively the best one to prescribe. The case against this kind of objectivism should get stronger as the thesis moves along, as I give more arguments for why normative reasons (and normative imperatives or 'oughts') are contingent on the psychology of the agent in question. Chapter 3 will contain more of these arguments.

If we're to be subjectivists about beliefs when it comes to normative reasons then there's no asymmetry for the reasons internalist to defend, an agent's reasons will completely depend on her moral psychology. But if my opponent doesn't want to be a subjectivist, and is persuaded by the strong intuitions about reasons that we have in certain cases like the gin example, then I suggested a way to find the middle-ground between the counter-examples to reasons subjectivism and the counter-examples to reasons objectivism. One method to determine whether normative reasons are based on the agent's beliefs or in the facts of the matter is, I argued, to think about persuadability: how easy it would be, hypothetically, to persuade the agent of the facts of the matter. This gave us plausible results in the most obvious cases and seemed to give the right kind of explanation; it was a way to track things that seem to be related to our concepts of normative reasons such as the risks involved and the availability of evidence.

Finally in **1.3** I introduced the most important distinction for my thesis overall: that of reasons internalism and reasons externalism. This thesis argues for the former: that all of an agent's normative reasons for action are contingent on her desires. I introduced some of the best arguments in the literature for reasons internalism, each of which ran along the lines of connecting an agent's normative reasons with actions she is capable (in a certain broad sense) of performing. I then discussed a particular objection that arises as a result of these arguments, to do with the belief objectivism / subjectivism about reasons that I discussed in **section 1.2**. I showed that the answers I gave in that section meant that the reasons internalist can justify the idea that reasons are sometimes contingent on the facts of the matter instead of beliefs (if they want to) without giving up the more important claim: that reasons are always contingent on an agent's desires.

The next chapter will continue to defend reasons internalism, this time against objections from Parfit.

Chapter 2.

What We Have Reason To Do:
A Defence of Value Subjectivism

Introduction

Chapter 2 will respond to the most serious remaining objections to reasons internalism, concentrating on arguments put forward by Parfit. Parfit argues against reasons internalism through two main arguments against 'value subjectivism', which I will explain and defend. This is the theory that value depends on the subject – which I'm willing to take as broadly as any creature with mental states – and their perspective, what *they* desire. It's worth noting from the beginning that there are at least two ways in which we can understand objective/subjective distinctions, as I showed in **1.2.1**: they can be taken, for example, as claims about relating to an agent's beliefs, and/or as relating to claims about an agent's desires. In this chapter I will only be concerned with value subjectivism in relation to the latter. This is because the relationship with desires is what's

important for the purposes of this thesis, but also (as I'll go on to show) because this seems to be what's important to Parfit as well.

In this introduction I'll set out what subjectivism is in more detail and then lay out the course for the rest of the chapter. Firstly, though, I'll briefly set out the discussion of subjectivism and criticisms of Parfit in terms of my thesis more generally.

Wider context

One of the aims of my thesis has been to determine when it is that an agent has a normative reason to act. Chapter 1 defends reasons internalism, in which any agent's normative reasons will be strictly related to bringing about what that agent desires.[60] Relating *value* to desire – as subjectivism about value does – is a similar project. After all, it seems intuitive that there's a connection between what reasons people have and what's valuable. By arguing that both value and reasons are related to desires, I will put together a picture of how this connection works. I'll also explore the relationship in more detail in future chapters, as well as looking at how normative reasons work with our moral obligations. More specifically, Chapters 3 and 4 will look at moral obligations and hypothetical/categorical imperatives, and build an account of when our obligations coincide with or come apart from what we have reason to do.

This chapter will support the work done in Chapter 1 both by helping to generally build up a picture of normativity and its relationship with desire and also, as I mentioned above, by directly defending reasons internalism against one of its main opponents. Parfit takes himself to be arguing against both value subjectivism *and* reasons internalism.[61] Because what's valuable is independent from our desires, he thinks, our reasons to act are similarly independent. So by defending value subjectivism from Parfit I'll also be defending my conception of internal, or 'subjective', reasons.

In the process of defending value subjectivism Chapter 2 will also try to better explain the motivations behind thinking that both value and reasons should be so *necessarily* linked with what subjects and agents desire. In Chapter 1 I discussed the importance of an agent's normative reasons being related to bringing about things that agent desires, drawing on work from those such as

[60] The details of that relation are unclear: it could be that the action actually has to have a chance of bringing about the desired outcome, or it could be just that the agent believes that it will. See Chapter 1 for details.
[61] Parfit, (2011a) p.54.

Goldman, Manne, Markovits and Williams.[62] In this chapter I will argue *against* a certain kind of intuition on why desire and value might be necessarily correlated: the intuition that pleasure comes from *desire-satisfaction*. Such pleasure, I will argue in **section 2.1**, is *not* what constitutes the necessary link between desire and value. This discussion will also ward off other objections to reasons internalism; Millgram, for example, argues against Williams' reasons internalism on the grounds of problems with a desire-satisfaction theory.[63] Millgram's objection is mistaken, I will show, because desire-satisfaction is not the motivation for a plausible account of reasons internalism. Indeed, I'll defend a conception of pleasure and pain as each being a kind of desire, and so not about desire-*satisfaction* at all.

What is subjectivism?

Onto defining subjectivism, then. Perhaps we desire things because we perceive them – the objects – to have value (and our perceptions may not be correct). Perhaps, instead, the things which are valuable are valuable because they are desired by subjects. This kind of question has been referred to as Aristotle's (and others') version of the Euthyphro question:[64] "*Do we desire things because they are good, or are they good because we desire them?*"[65] The debate between objectivism (roughly the former option) and subjectivism (roughly the latter) about value is one that tries to determine the answer to that question, and/or resolve differences between those positions. This chapter will defend subjectivism and agree that what's valuable is valuable in virtue of its being desired by subjects. Although I won't put forward any new positive arguments in favour of the position,[66] I will defend it against what I take to be the best and most prominent counter-arguments.

I'll briefly say something more explicit about the relation between reasons internalism, subjectivism about reasons and subjectivism about value, and try to briefly map out the terminology. As Parfit describes them, "Subjective theories appeal to facts about our present

[62] Williams (1981), Goldman (2009), Manne, (2014), Markovits, (2014).
[63] Millgram, (1996) p.206.
[64] Aydede calls it the Euthyphro problem in Aydede (forthcoming a) and Street refers to it as a "modern, secular version" in Street, (2009) p.274.
[65] Finlay, (2014) p.4, for example, takes Aristotle to be asking this question in The Metaphysics. Finlay also notes that it's related to the wider question that Smith termed 'The Moral Problem' in Smith, (1994) of how to reconcile morality's objectivity with its normativity. This is a topic that I'll refer back to more specifically in later chapters.
[66] I'll be defending subjectivism only in terms of Parfit's own arguments against it. For other arguments for subjectivism see, for example, Korsgaard, (1996b) and Goldman, (2009).

desires, aims and choices".[67] Reasons internalism is a *subjective* theory, because it is a thesis which links agents' normative reasons to their present set of desires. Confusingly, for my purposes, in On What Matters (particularly Volume 1), Parfit tends to argue against what he calls 'reasons subjectivism' instead of reasons internalism. For the rest of my thesis I'll take these to be the same theory under different names, but I'll continue to refer to it as 'reasons internalism', after Williams. In this chapter, then, and unless otherwise specified, when I talk about 'subjectivism' I mean 'subjectivism about what's valuable'.

Subjectivism about reasons is a different thesis to subjectivism about what's valuable because, as the names suggest, one is a theory that links desires with reasons and one is a theory that links desires with what's valuable.[68] Parfit argues against reasons internalism and value subjectivism at the same time, since he has a 'buck-passing' view of value,[69] that the goodness of an object is constituted by other properties which can give us favourable reasons to act in certain ways towards it.[70] I won't have time to discuss the buck-passing account in much detail here, but it should be sufficient to explain that to Parfit value and reasons are linked in such a way that to argue that value cannot be necessarily connected with desires is the same as arguing that reasons don't have that connection either. Again, when I refer to subjectivism in the rest of this chapter I'll be referring to the position that what's valuable is contingent on what agents desire.

In understanding the subjectivist's position, it's important to note the differences between subjectivism and other theories that it might be confused with. Here I'll take a quick detour to discuss the difference between value subjectivism and moral relativism. Moral relativism is "the thesis that the truth or justification of moral judgments is not absolute".[71] According to moral relativism, there's no absolute or universal truth as to what's morally good, because it varies between situations, people or places. To argue that value is *subjective* is importantly different from arguing that morality is relative, and even if morality is based on what's valuable then moral relativism doesn't necessarily follow. It's coherent to still have an absolute and real morality based on a subjective theory of value. Moral theories can be objectively true even if they promote / honour / bring about (respond in whichever way(s) it is appropriate for a moral theory to respond to) value that is dependent on subjects and their desires. This is because the subjective nature of

---

[67] Parfit, (2011a) p.58. This supports what I said at the beginning of this chapter: that, for Parfit, desire is the main concern here, rather than belief.
[68] Goldman also explains the relationship between internal reasons and value subjectivism in Goldman (2009) p.11.
[69] The buck-passing account of value that Parfit adopts is originally from Scanlon, (1998).
[70] Parfit, (2011a) p.38. For other discussions of the buck-passing account see, for example, Heuer (2010), Gregory, (2014) and Skorupski, (2007b).
[71] Gowans, (2016).

the value doesn't affect whether moral judgments are true or justified. It could be true that it's good to promote whatever it is a certain subject values, for example. (Although obviously, depending on whether what that person values clashes with what other people value, that act might have bad features for someone as well as good ones for someone else.)

Subjectivism is also not the theory that *anything* can be valuable. This would only ever be the case to the extent that subjects could truly desire anything, and *that* may not be the case. Furthermore, subjectivism isn't committed to *all* desired things being valuable. It might be the case that only certain kinds of desire are indicative of what's valuable. I'll discuss this in greater detail in **section 2**, when I argue that pleasures and pains are pleasurable and painful in virtue of the subject having a specific kind of desire. In Chapter 3 I'll go into detail about why the subjectivist theories I defend in this thesis are compatible with moral realism.

This chapter

As I've said above, this chapter will defend subjectivism specifically from Parfit, who simultaneously argues against both it and reasons internalism. **Section 2.1** will examine the different approaches he takes, categorised into two main types: his argument on 'hedonic likings' and his argument based on future desires, known more commonly as his 'agony argument'. Next, **section 2.2** will address the former argument, and defend against it by arguing in favour of the 'desire account' of pleasure and pain. This is the thesis that pleasure and pain (and therefore what Parfit calls 'hedonic likings') are subjective, and what it is for an experience to be pleasant or painful is for the subject to have a specific kind of desire.

Section 2.3 will address the agony argument and discuss whether future desires are valuable in the present. I will conclude that they are not, and that Parfit's examples to the contrary are not as intuitively plausible as he makes them out to be.

2.1 Parfit's Arguments against Subjectivism

2.1.1 Introduction

Parfit has two main arguments against subjectivism. In this section I will introduce them and separate them into two main arguments: one on hedonic likings and one on future desires.

2.1.2 Hedonic likings

When philosophers argue that value is subjective, that things are valuable because *we desire* them, Parfit attributes this to a mistake about what kinds of states, or which parts of mental states, can make something valuable.[72] The mistake is between two distinct parts of mental states: 'hedonic likings' and 'meta-hedonic' desires.

Hedonic likings and dislikings are something like the goodness or badness of sensations that we feel when something is pleasant or unpleasant. Parfit describes them thus:

> [An] important set of mental states, though they are often assumed to be desires, are better regarded as being in a separate category. These are the hedonic likings and dislikings of certain actual present sensations that make our having these sensations pleasant, painful, or in other ways unpleasant, or in which their pleasantness or unpleasantness partly consists.[73]

Parfit lists hunger, thirst and lust as examples. He makes sure to distinguish hedonic likings and dislikings from the bare sensations themselves:

> Though these [bare] sensations are not in themselves good or bad, they are parts of complex mental states that are good or bad. When we are in pain, what is bad is not our sensation but our conscious state of having a sensation that we dislike. If we didn't dislike this sensation, our conscious state would not be bad.[74]

So far we have at least two distinct and separate parts of the mental states we're in when we find experiences pleasant or unpleasant: the physical feeling and the hedonic liking or disliking.

---

[72] Parfit, (2011a) p.55 refers explicitly to Korsgaard here as someone who makes this mistake about what's valuable.
[73] Parfit, (2011a) p.53.
[74] Parfit, (2011a) p.54.

Finally, we come across a third distinct part of the experience which Parfit refers to as the meta-hedonic desires:

> When we are having some sensation that we intensely like or dislike, most of us also strongly want to be, or not to be, in this conscious state. Such desires about such conscious states we can call meta-hedonic. Many people fail to distinguish between hedonic likings or dislikings and such meta-hedonic desires. But these mental states differ in several ways. What we dislike is some sensation. What we want is not to be having a sensation that we dislike.[75]

Hedonic likings can "confer value",[76] but meta-hedonic ones cannot. This is the crux of Parfit's 'hedonic likings' argument against subjectivism. The part of pleasant or unpleasant sensations that makes them valuable is separate to the part which constitutes our desires about the situation. Value is therefore not reliant on desires, and not subjective.

2.1.3 Agony argument

A second argument of Parfit's against subjectivism is the agony argument.[77] According to this argument, to a subject who has no desire or aim to avoid agony, there would be nothing disvaluable (on a subjectivist account) about such agony while it's still in the future. Because such a conclusion seems so implausible then subjectivism must be wrong.[78] Sobel describes the argument clearly when he says,

> The first premise of the Agony Argument is that we have current reasons to avoid future agony. Its second premise is that subjective accounts cannot vindicate this fact. So, the argument concludes, subjective accounts must be rejected.[79]

Sobel talks in terms of reasons, instead of in terms of value, but for Parfit it is an argument against both value subjectivism and reasons internalism. Just as subjectivism about value would not be

[75] Parfit, (2011a) p.54.
[76] Parfit, (2011a) p.55.
[77] As named by Parfit in Parfit, (2011) p.73.
[78] Parfit (2011a) p.74.
[79] Sobel, (2011a) p.52.

able to ascribe disvalue to this person's future agony, so reasons internalism would not be able to demonstrate why that agent has reason to avoid it. I'll largely stick to the language of value where I can in this section for the sake of consistency, although some of the quotes I use from other sources will focus on reasons instead.

Parfit's argument here is different to the previous one, because he's not trying to argue that the subject wouldn't necessarily have desires at the time about avoiding pain. Instead, he wants to make claims about "the difference between our attitudes to present and future agony."[80] He doesn't think that a theory of value that relates necessarily to an agent's desires can be a reliable way for the agent to get what actually ends up being good for them. It's an understandable concern.

The argument also appears in his book *Reasons and Persons* (p.124). He invites us to picture a subject who has no desire to avoid agony on *future* Tuesdays, at least not until the Tuesday arrives:

> It might ... be claimed that my predictable future desire to not be in agony gives me a desire-based reason now to want to avoid this future agony. But this claim cannot be made by those who accept subjective theories of the kind that we are considering. These people do not claim, and given their other assumptions they could not claim, that facts about our *future* desires give us reasons.[81]

If the agent doesn't have the desires now, then Parfit argues that the subjectivist cannot appeal to those desires in explaining what's valuable or what they have reason to do.

It's also worth mentioning how widely applicable this criticism might be. Although a person blind to future agony on Tuesdays is a bit far-fetched, my opponent might worry that there are many real-world situations in which this happens, even if it might be less extreme. Looking after one's short-term happiness is often a lot easier than looking after, for example, one's long-term health: a phenomenon Parfit calls 'bias towards the near.'[82] Broome and Parfit also discuss a similar example in the co-authored paper 'Reasons and Motivation', in which they argue that someone who will suffer in the future otherwise has reason to take medicine today, regardless of whether they have any current desires to avoid suffering in the future.[83]

[80] Parfit, (2011a) p.74.
[81] Parfit, (2011) p.74.
[82] Parfit, (1984) p.124.
[83] Broome and Parfit, (1997).

Having established two different arguments that Parfit makes against subjectivism, the rest of this Chapter will turn to the responses. **Section 2.2** will argue that Parfit is wrong to distinguish between hedonic and meta-hedonic desires and **2.3** will follow Street and argue that cases like the Future-Tuesday case are so alien that the conclusion is not so difficult for the subjectivist to accept after all.

2.2 Pleasure and Pain as Subjective

2.2.1 Introduction

This section will address the 'hedonic likings' argument that Parfit uses to argue against subjectivism. Recall: Parfit argued that when it came to certain mental states that might be valuable or disvaluable, such as pleasure and pain, the part of the mental state that 'confers' such value is not the same part, nor is it necessarily connected to, the desire the agent might have about that mental state. For example, a desire for pain to stop is not the disvaluable part of pain, rather it is the 'hedonic liking' itself. This distinction doesn't work, and in this section I'll show why by arguing that the part of mental states which makes them pleasurable – what Parfit refers to as the 'hedonic likings and dislikings' – is the very same part, and cannot be separated from, the subject's desires. I will do this by defending what's known as the 'motivational account' of pleasurable and painful experiences, but that (to pre-empt an objection) I will come to call the 'desire account'.

My discussion will be split into several sections. I'll begin by explaining the desire account in detail, including discussions about the kinds of pleasurable and painful experiences and the kinds of desires it covers. I will clarify that it is not an account of desire-satisfaction but an account in which pleasure and pain *are* a kind of desire. I will also clarify that the account isn't limited to only pleasurable and painful sensations that are *physical*. Turning to pain specifically, I will clarify the differences between experiences that are painful and experiences that are simply unpleasant.

In the next section I'll list two main arguments for the desire account: the heterogeneity argument and the consistency argument. The former argument has been made elsewhere, but the latter, I will argue, is all that's needed for the thesis to be plausible. I will make the consistency argument: demonstrating that all cases of pleasurable and painful experiences do correlate with a certain kind of desire, and taking on a wide-variety of prominent counter-examples along the way.

Hopefully the scope of the counter-examples I reject will be wide enough to make a generally convincing case.

2.2.2 The motivational / desire account

Chris Heathwood worked on a precise formulation of the motivational account of pleasure, as he called it.[84] My version of the thesis, largely taken from his, is this:

A subject is having a pleasant experience at any given time T if, and only if, at T they non-derivatively desire for that experience to continue. The converse applies for unpleasant experiences.

What it is for an experience to *be* a pleasant or an unpleasant one for the subject is, according to this thesis, that the subject experiences a non-derivative desire for that experience to continue, at the time that it is happening, where a non-derivative desire is one that is desired for its own sake, as opposed to being desired instrumentally: *only* in order to further other aims.[85] By the 'experience' I mean something closer to what Parfit meant when he spoke about the 'bare sensations', which refer to the way that the experience appears in the mind of the subject, rather than the 'hedonic likings'. I'll clarify what I mean by this part in more detail in the second subsection.

Although Heathwood talks of the 'motivational account', I will call it the 'desire account'. This is to pre-emptively ward off some of the criticisms that will come up later in the section. After all, the thesis is that the pleasantness or unpleasantness is constituted by a desire, not by actual motivation. It's not the case that unpleasantness or pleasantness *will* always correspond to motivation, let alone be constituted by it. But I take this to be a terminological mistake on Heathwood's part that has led to some misunderstandings, rather than a difference in the substance of our theses.

The desire account is more than just a claim about whether the desire and the pleasantness happen to occur at the same time: it's a thesis that the pleasantness or unpleasantness of the

[84] Heathwood, (2007). He followed and developed upon previous work on this topic by authors such as Alston (1967), Brandt (1979), Carson (2000), Korsgaard (1996a) and Parfit (1984).
[85] For more discussion on the difference between instrumental and intrinsic desires see, e.g., M. Schroeder (2004) p.181 and T. Schroeder (2017).

experience is *constituted by* the desire. Mark Schroeder helpfully describes a constitutive relationship when he says,

> Constitutive explanations, I take it, are a ubiquitous phenomenon with which we need to be comfortable in order to understand a wide variety of phenomena. Figures are triangles by having three sides; they are not three-sided by being triangles.[86]

This is the kind of relationship the desire account is arguing for. An experience is pleasurable by the agent having the right kind of desire.

I'll next mention a few things that the desire account *is not*, in the hopes of making it even clearer what it *is*. Firstly, I am not advocating the idea that a subject's desire to stop having an unpleasant experience would always be overriding, or that they must necessarily desire to stop having that experience more than they would like to continue it. I am happy to claim that agents can have conflicting desires; my desire to have a nap is not silenced by my desire to continue writing an essay, my desire to have caramel ice-cream is still strong in spite of my desire to have mint choc-chip, and even though I think they'd taste bad when put together.

I will also claim that a subject can have multiple different experiences at one time. The experience or part of the experience that the subject wants to stop is the part that they find unpleasant.[87] At a given time, for instance, a subject could be experiencing both a sharp pain from a needle in their arm and a certain satisfaction from knowing that they are donating blood. The fact that the latter is enjoyable wouldn't necessarily stop the former being unpleasant; they may at the same time want the positive feeling to continue while being keen for the sharp pain to end. And although they may desire to experience the sharp pain, they do so not *intrinsically* as specified above, but rather instrumentally: to do something good.[88]

The desire account is also distinct from, and doesn't either entail or follow from, hedonism. As Moore writes in the *SEP*, listing two different kinds of hedonism,

---

[86] M. Schroeder, (2004) p.63.

[87] I take it that an experience can be divided into several parts, which are also experiences. For example, I can have the experience of eating ice cream as a whole and I can also have an experience of the coldness of eating ice cream.

[88] Kagan has a similar discussion on whether we can ever have good reason to desire pain despite its disvalue. He says, "Intrinsic disvalue does not rule out the possibility of extrinsic value", and so there might always be positive things that we can bring about from things which, on their own, are intrinsically negative. Kagan, (1989) p.168.

> Psychological or motivational hedonism claims that only pleasure or pain motivates us. Ethical or evaluative hedonism claims that only pleasure has worth or value and only pain or displeasure has disvalue or the opposite of worth.[89]

Unlike hedonism, the desire account doesn't attempt to account for everything that's valuable or for what motivates agents. It's simply an account of what it means for an experience to be pleasant or unpleasant to a subject. A positive argument in favour of pleasure and pain being valuable is beyond the scope of this chapter, but it's also not something that is necessary to defend subjectivism and reasons internalism against Parfit's arguments. Parfit's arguments, after all, are that the subjectivist cannot properly account for the value of pleasure and pain, not that they aren't valuable to begin with.

The rest of **2.2.2** will now go through some more detailed clarifications. Issues that I *won't* discuss include those of self-knowledge and self-ignorance. I don't aim to settle the question of whether the subjects will always know or believe that their experiences are pleasant. But I take it that this can be just as much a question about the transparency of the agent's desires as it is about whether the experience is pleasant.

On an account of desire-satisfaction

Another important clarification is that the desire account is not an account of pleasurable/painful experiences as one that explains them in terms of desire *satisfaction*. The desire account, after all, is an attempt to describe a certain link between what agents desire and what they find pleasurable, and one such phenomenon that one might want a thesis to describe is why people derive pleasure from their desires being satisfied. But it's not a necessary link and there are many examples where desires' satisfaction might not result in pleasure for the agent.[90] Some literature that criticises the desire account does so because it mistakes the desire account for an account about desire-satisfaction, perhaps because the opponents incorrectly see the desire account as trying to explain that kind of link between desire-satisfaction and pleasure.

---

[89] Moore, (2013).

[90] Such an account (one that explains pleasurableness in terms of desire-satisfaction) has been satisfactorily rejected elsewhere, eg Plato's *Philebus*, (2000), Brandt, (1992), and Katz, (2016). Katz, as I'll show, specifically mistakes the desire account for an account of desire-satisfaction.

Katz, for example, describes Heathwood's desire account as saying that "pleasure is definable as believed satisfaction of current desire"[91] and then gives criticisms that would only work against an account that does rely on the satisfaction of desires (despite going on to claim that he has refuted satisfaction of desire *and* desire itself as a possible criterion for pleasantness/unpleasantness).[92] But the desire account is silent on desire-satisfaction, despite Katz's claim. Subjects who desire for their experiences to continue may well have their desires satisfied as the experience continues, but that's not what makes the experiences pleasurable according to the account. It is the *current* desires which constitute pleasure or pain.

There's one more point I'll briefly address before I move on to the next general clarification. Some positive experiences, for example, we'd not want to experience indefinitely. But the desire for an experience to continue is more immediate than a desire about the future generally. Just like the desire for a pain to stop is a desire for it to stop instantly, the desire for an experience to continue is only a desire for it to immediately continue, rather than a desire for it to continue for any particular length of time. In fact, an easier way to understand the desire might be as a present-directed one instead of a future-directed one. It's a desire that the experience the subject is having is an experience which is continuing.[93] I might, for example, be having a lovely time having coffee with my friend, and have a desire for *this* experience, the bare sensations that I am having, to continue. But that isn't the same thing as wanting to spend the rest of the day having coffee here with my friend. After all, we both might get a little annoying if we spend too much time together, and I get over-caffeinated fairly quickly anyway.

On whether sensations are physical

Heathwood also discusses the difference between sensory pleasure and other varieties.[94] He says, to describe cases of sensory pleasure, "it seems clear that there are sensations, or feelings, of pleasure. If you're like me, you continually experience sensations, and some such sensations you would not hesitate to describe as pleasant."[95] This is in contrast to the ways in which we say we

[91] Katz, (2016).
[92] "So it appears that it won't do to make either desire or its satisfaction or sensings or beliefs in that satisfaction sufficient for pleasure, let alone identical to it, as these philosophers have variously proposed." Katz, (2016)
[93] There are similarities between the desire-account and other 'attitudinal' accounts, which might be based on other kinds of attitudes that the subject has rather than desires. But I take the desire-account to be better, on the grounds of being more informative.
[94] Heathwood, (2007) p.28.
[95] Heathwood, (2007) p.28.

are pleased about things without actually experiencing any pleasurable sensations. Being pleased about the winner of a game, for example, or about the state of the economy.

I will follow Heathwood's definition to an extent – and the reason it's limited to a certain extent is more to do with my not understanding his distinctions than my necessarily disagreeing with him. There are, as there are with Heathwood, several phenomena that I wouldn't want to include in my discussions. Perhaps, for example, something just makes a subject feel *less bad*, rather than *more good*, and I wouldn't want to count that as a pleasant experience. Perhaps something that a subject desired to happen comes true and they might call it pleasing while not actually experiencing pleasure. In each of these cases we might describe ourselves as *finding* pleasure in something while not having a sensational experience of it. But I might be less strict than Heathwood in other ways, and I will take a moment now to clarify how in a bit more detail. I don't want to restrict my discussion only to the kinds of pleasant experience that produce a *physical* sensation, for example, and I'm not sure about the extent to which this is the kind that Heathwood himself wants to refer to; he compares sensory pleasure to 'propositional pleasure' and 'enjoyment' but I find the distinctions murky.

Aydede also tries to flesh out the distinction, and compares sensory pleasure to a pleasure he describes as "more cognitive, (conceptual, higher, intellectual, etc.)"[96] He asks us to imagine the difference between two ways of having a pleasant experience of eating watermelon: firstly in enjoying the taste simply because one finds it tasty, secondly in enjoying the taste because (despite not being fond of the taste in the usual way) the taste indicates that the watermelon crop is objectively a good one, and the taster enjoys the taste because of other facts that the taste entails.

I don't see a reason to restrict the account in that way, particularly since the bigger picture of this chapter concerns what's valuable, and it's in my interests to make the kind of pleasure I'm talking about as broad as I can while still being pleasure. I will understand pleasurable experiences to be those experiences that the subject experiences *as pleasant*, whether or not their experience produces a physical sensation or feeling (like the more mechanical sensation of watermelon tasting), and whether or not there is something more conceptual about the experience. As long as it's the experience itself that's pleasant, which it seems to be in both kinds of case, I don't think that further detail is needed. Indeed, the desire account is itself an attempt to clarify and explain these experiences in more detail, so any further attempt to get into the details might be too much like skipping ahead.

[96] Aydede, (forthcoming a) p.4.

On pain versus unpleasant experiences

I said above that the desire account applied both to pleasure and to pain. To call the opposite of pleasure 'pain' is problematic, because pain itself is a tricky phenomenon with extra philosophical baggage. It would be more accurate to refer to experiences that are unpleasant. The two obviously have significant overlap: unpleasant experiences are often painful, and pain tends to be an unpleasant experience. Here I'll briefly discuss examples where pains are supposedly not unpleasant experiences and when unpleasant experiences are not painful.[97]

Firstly, there are times when subjects have been described as experiencing pain but *not* finding those experiences unpleasant. I take this supposed discrepancy to be down to a distinction between the physical sensations of pain and pain of a different description. The discrepancies will usually be down to something anomalous in the body's functioning such as pain asymbolia, a lobotomy or taking morphine.[98] Aydede describes this distinction thus:

> There are two main threads in the common-sense conception of pain that pull in opposite directions. We might call this tension the act-object duality (or ambiguity) embedded in our ordinary concept of pain.[99]

As he goes on to say, the first thread is seeing pain as something in a body part, the second is seeing it as a subjective experience. When subjects are described as being in pain but not finding the experience unpleasant, it seems to be pain in the first sense. The subjects might experience the same physical sensations as they would if they were in pain normally, but without the subjective part of the experience which determines whether the experience is unpleasant.

For my purposes, I am only interested in pain insofar as it is an unpleasant experience, so including the subjective part of the definition. Scientific consensus is that this is a good definition for pain. Aydede quotes the definition given by the International Association for the Study of Pain (IASP),

[97] See also Corns, (2014).
[98] For discussion see Aydede, M. (2013). Street also discusses these cases in Street, (2005) p.147-148.
[99] Aydede, (2013).

Pain: An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.[100]

It also seems to be the most ethically relevant kind of pain. Any reason why pain is of intrinsic negative value will, after all, be because of the unpleasantness of pain. Experiences physically similar to pain but that aren't unpleasant don't seem to be the kinds of states we would have a reason to avoid except for instrumental reasons (such as an underlying health problem).

The second kind of example of when pain and unpleasant experiences don't correlate is when subjects can have unpleasant experiences that aren't painful. These seem to be both more common and less controversial. Having a philosophy paper rejected from a conference will often be an unpleasant experience but won't always be *painful*, as such. Accidentally putting one's foot into something sticky could also be pretty unpleasant without being painful.

The IASP's definition also gives a satisfactory answer as to when an unpleasant experience is also a painful one, and that's when the unpleasant experience is associated with or described in terms of tissue damage. This seems to match common-sense intuitions about when something is painful rather than *just* unpleasant.

I've briefly discussed the relationship between pain and unpleasant experiences. As per the consensus of the scientific community, pain is always an unpleasant experiences and unpleasant experiences are sometimes painful, when certain other conditions are met. For my purposes, then, since I am interested in unpleasant experiences, I am also interested in painful ones.

On desire

An account of pleasant experiences which bases them on desires can be fairly neutral on what desire itself is. And as I argued in the introduction to the thesis, I want to make my arguments as compatible as I can with different accounts of desires. But I cannot be entirely neutral; if pleasures and pains are each defined as a certain kind of desire then desires can't themselves be defined as pleasures and pains – not without getting trapped in a viciously circular explanation. By arguing in favour of the desire account of pleasure I am also arguing against those particular theories of desire, but beyond that I won't take any sides in the debate. Since there are plenty of other theories,

[100] Aydede, (2013) and for a defence of the definition see Aydede, (forthcoming b).

and ones that reduces desire to pleasure are by no means the most popular, this shouldn't be a controversial move to make.[101]

Being flexible about what it is to desire is still consistent with the overall aim of the chapter. The main reason I'm focusing on the desire account is to demonstrate that the pleasantness or painfulness of an experience is subjective – that it depends on the agent's attitudes – rather than objective. This isn't something that I need a particular theory of desire for; any should do.

2.2.3 Two arguments in favour of the desire account

The desire account of pleasurable experience is an attempt to explain what it is about a pleasurable experience which actually makes it pleasurable. There are two main arguments for why an agent's desires are the best answer. Firstly there's an argument of consistency: whenever there is a feeling of pleasure there is also a desire for it to continue, and vice versa. Whenever an experience is unpleasant, there is a desire for that experience to stop. Indeed, I find the desire for an experience to stop to be the most prominent part of pain, and one that increases in proportion with how unpleasant the experience is. The more intense the pleasure, the stronger the desire for it to continue.

The second argument for the desire account of pleasure is that the relevant desires for the pleasure to continue or pain to stop are the only part of those experiences which remains the same across different kinds of those experience. The sensations of ice cream on the tongue or of music to the ears are both experienced very differently, except in the way that the subject (assuming she finds them both pleasurable) desires for the sensations to continue. This is the 'heterogeneity problem',[102] levelled against other accounts of pleasure. Other qualities of pleasurable experiences vary so much that the only thing which brings the examples together, according to the argument, is the desire.

The heterogeneity problem has been discussed elsewhere as an argument against other accounts.[103] My own work to defend the account will be through the argument of consistency. Demonstrating an homogeneity in this way is obviously a difficult task because it's impossible to cover every possible instance of a pleasurable experience or of the relevant kind of desire. I'll

[101] I briefly list some of the alternatives in the introduction to this thesis.
[102] See, for example, Aydede, (forthcoming a) p.7-8, Heathwood, (2007) p.26, Korsgaard, (1996a) p.148 and Feldman, (2006) p.79.
[103] For example, Heathwood, (2007) p.25-26

overcome this challenge in two steps: firstly I'll briefly discuss some fairly everyday examples of pleasurable and painful experiences of different intensities, and demonstrate how the pleasurableness or painfulness of the experiences seem to match up with the desires. I will then introduce and refute the most prominent and contemporary purported counter-examples to the desire account and indicate what kinds of mistakes these counter-examples make. This approach will then hopefully be widely applicable to other counter-examples that my opponents might consider. By approaching any counter-examples by their type and refuting them in a methodical and thorough manner, I hope to make a persuasive case that *all* examples of pleasure and pain correspond with the relevant kinds of desire and that to have a pleasurable or painful experience *is the same* as having those kinds of desires. I'll have therefore demonstrated that pleasure and pain are *subjective*.

Consistency: strong and vivid experiences

Now, I'll show in three stages how an account of someone feeling pleasures and pains is consistent with their intrinsic desires about those situations, even across differing intensities. My aim here is to help explain what it is that I find really motivates a belief in the desire account. It's important to understand what can motivate the theory if we are to understand the theory itself. This is at least partly because I want to move away from false conceptions of desire theories, such as the conception of it as something that tries to explain pleasurableness in terms of desire-satisfaction, as I rejected in **2.2.2**. I want to make it clear that the desire account doesn't rest on any intuitions about satisfying our desires being a thing we find pleasurable. The desire account is plausible, instead, on grounds of consistency. I'll briefly describe three stages of a subject coming across an intense pain, and going through how in each case the intensity of pain correlates with the intensity of desire.

When I think about intense pains what strikes me most is a desire for the pain to go away.[104] This, obviously, is a good starting place to find some evidence for the desire account of experience. Suppose – rather embarrassingly – a person falls off a treadmill. The experience can be described, step by step, in terms of how much the person desires for their experience to stop and how much pain they're in. Take the following stages of a painful experience:

[104] Manne also has an excellent description of certain experiences – including pains – as being a kind of 'make-it-stop' state for the subject. She discusses it in Manne (forthcoming).

1. ‘L’ tumbles off the treadmill, banging several parts of her body.

As she falls she doesn’t have much time to realise what’s happened, and she doesn’t immediately have a desire for the experience to stop; it’s all happened too quickly. The initial experience is just awareness of falling, nothing subjective. But this isn’t a problem for the thesis because just as we can say she doesn’t yet desire for it to stop, it also seems to be the case that she isn’t yet feeling any pain. It’s all happened too quickly to be much aware of the physical impacts and the overriding feeling is more that of confusion and perhaps a rush of adrenaline than it is of pain.

2. ‘L’ lands on the floor and comes to a stop.

After the immediate rush wears off, the subject now finds herself in a fair amount of pain. She can feel multiple bruised bones and stinging where she’s grazed her knees. Here is when she feels the most pain and – correspondingly – the strongest distress and desire for it to get better.

3. ‘L’ waits for the pain to subdue.

Having acknowledged and accepted her fate, our protagonist now needs to wait out the immediate pains until she can get up and go home. As she does so she tries to make herself feel better by distracting herself with her breathing and filling her mind with other thoughts. As she does so, she feels less strongly the desire for the pain to go away. But, correspondingly, she doesn’t feel the same amount of pain; although the same physical symptoms are there her distraction makes it at once less unpleasant, less distressing, and less painful.

Mild pleasures and pains

The strength of pains or pleasures correspond with the strengths of the relevant desires. The milder experiences of pleasures and pains are all accompanied by equally milder desires. Where a very strong hunger is very painful and comes with a very strong desire for it to stop, a milder hunger is often barely noticeable and barely painful at all. To the extent that it is a desire, it is a pain, and vice-versa.

It's worth noting that the desire is for the experience itself – the sensory experience – to continue or stop. There might be other desires, even ones that tend to go hand in hand with the experiences, that aren't directed towards those experiences. Those desires *won't* necessarily correlate with the intensity of the experience. In the case of hunger, such a feeling will often be accompanied by the desire to actually go and get food, but sometimes it won't. It's only the intrinsic desire to not feel hungry which is what necessarily corresponds to how pleasant or unpleasant it is.

There are also lots of mild pleasures in life. The gentle feeling of a breeze or some sunrays, for example. As long as it's a pleasure then that feeling is accompanied with a desire, even a mild one, for it to continue. I'll use these examples to explain another reason why it might seem that a subject would sometimes not desire for a pleasant experience to continue. Desires often don't last, and desires for milder pleasures seem like a good example of this. After a while we might want to get away from the sun, or find ourselves wanting to shelter from the breeze.[105] But again, this corresponds to how we feel and how the pleasure we might get from the experiences will change: the sunrays and the breeze won't always continue to feel pleasant and our desires will diminish accordingly.

Feeling no pleasure or pain

When there's no sensory feeling of pleasure or pain then there are no intrinsic desires, as described above, for those feelings to continue or stop. The converse is also true. A feeling of perfect contentment might be pleasurable – in which case the subject will desire for it to continue – or it might sometimes just be a neutral state. In the latter case a subject might feel many things, including many kinds of desires, but none of them of the intrinsic kind specified by the desire account.

Summary

Hopefully this section has done more than just methodically list different levels of pleasure and pain and what their corresponding desires would be. I hope to have also started to make the desire account plausible. I've shown how a variety of cases of pleasures and pains can be equally described in terms of the relevant desires. To add to my evidence, the next sections will go through the most

[105] This pre-empts a criticism from Goldman in Goldman, (2009) p.230 where he worries about whether the desire thesis can still account for pleasures that we don't want to last very long, because of the kind of pleasures that they are. Eating ice cream, for example, isn't something you'd want to do for the rest of your life. The desire thesis is safe, though, because the feeling of eating ice cream simply won't always continue to feel pleasant just as the desire for the experience to continue won't last.

controversial examples I can find and I'll draw attention to some common mistakes that might be made in the process. Since I'm arguing for the desire account with an argument of consistency, defending the thesis against these counter-examples should also serve to strengthen my argument, and make it even more plausible that *all* cases of pleasant or unpleasant experiences have the relevant, corresponding desires.

2.2.4 Reflective blindness objections

Bramble refers to the 'Reflective Blindness' objection as a 'decisive' objection against the desire account, as well as against any other account which relies on the subjects' attitudes. [106] Reflective blindness, here, is a phenomenon in which the subject is supposedly having a pleasant (or unpleasant) experience, but isn't aware of it at the time. If the subject isn't aware of it, according to Bramble, they cannot have the relevant desires that would explain (on the desire account) why the experience is pleasant/unpleasant.

I'll give two examples of reflective blindness and then go on to show that one of two things is happening, depending on the specifics of the examples: either these examples are incorrectly categorised and the subject is aware of the experience but not of certain details, or the examples aren't of the subject having a pleasant or unpleasant experience at all.

Example one – depression

The first example of reflective blindness is depression: it can often affect mood and attitude slowly and gradually, without the subject being aware. The subject may never have thought to categorise themselves as having depression. Rachels uses this as an example,[107] which I'll talk more about specifically later, and it's also mentioned in an excerpt of Haybron that Bramble uses when discussing reflective blindness.[108]

Example two – old age

The second example I'll discuss is the physical pains and aches that can creep up on someone as they age. This comes from Bramble:

---

[106] Bramble (2013).
[107] Rachels, (2000).
[108] Haybron, (2008) p.222 also quoted in Bramble, (2013) p.205.

> For a [clear] example, imagine being suddenly transported into a younger body. Isn't it likely you would learn immediately, due to the contrast, of unpleasant experiences you had been having in your older body that you had been completely unaware of at the time (say, ones due to physical pressures being put on your body as a result of aging)? Unpleasant experiences seem to be capable of sneaking up on us by starting in very small amounts or very low intensities and then slowly accumulating or intensifying over time. In this way, we can come to suffer a considerable amount without ever having any idea of it. Unpleasant experience of this variety we might refer to as 'suffering by stealth'.[109]

These unpleasant experiences come about so gradually that the subjects have no idea. There's no one moment when suddenly they're in pain, it's just something which develops in miniscule amounts but builds up to something more significant.

A response to cases of reflective blindness

Each of the examples that Bramble provides, including the two that I've listed here, make one of two mistakes about the desire account. Either the subjects do have the relevant kinds of desire, but they're directed at a different part of the experience to the part that the agent is unaware of, or the cases aren't examples of unpleasant experiences at all.

Depression – response one

Depression is, as Bramble is right to imply, a good example of when a subject can be unaware of the kinds of unpleasantness that they're experiencing. But although these subjects may not know that they're depressed, they can still be aware of other parts of their experience. Suppose we take a subject who has never yet thought of herself as depressed, but has gradually become so over a few years. We would correctly describe depression as the cause of her suffering, as it causes her everyday experiences to be difficult, wears her out, takes the pleasure away from the things she used to enjoy. To the sufferer, 'being depressed' is not necessarily what she finds unpleasant (since she does not know she's depressed): it's the experiences that depression affects that she finds unpleasant. Similarly, it is not necessarily the depression that she intrinsically desires not to be experiencing (for the above reasons) but the everyday experiences that the depression is affecting:

[109] Bramble, (2013) p.206.

plausibly, she can have a desire not to have the experiences of going to work, going out to the shops, even waking up, without ever having noticed that she has depression.

Perhaps critics would not want to agree that the depressed agent's displeasure can really be said to come from the mundane experiences of going to work or to the shops. It might be argued that such a list of causes, one that doesn't include depression itself, would be inadequate: not really touching on the source of the agent's unhappiness or accurately accounting for it. After all, going to the shops seems like a harmless experience, but to have depression is a horrible one. The latter is the cause of all of the unpleasantness, not anything to do with the trip outside. To say that the agent's unpleasant experiences only come from the culmination of individual bad experiences might seem like an incomplete description.

After a closer look, though, this is not a problem for my account. Everything that could be said to be unpleasant about depression comes from the way the subject experiences life normally, so these experiences are also the ones that the subject desires to avoid, and so also the ones that make her circumstances unpleasant. Neither does my position downplay the severity of her condition; her depression is no less terrible by being described this way. Every reason why depression makes her life unbearable can be described in terms of what she experiences and the way depression affects those experiences. This is the case no matter how the subject is able to characterise what is happening to her; even if she has no idea that she is depressed, or that the cause of her unhappiness is a medical condition, she will still have that intrinsic desire to avoid whatever experiences she finds unpleasant.

Depression – response two

To the second part of my response. Suppose we consider a subject who doesn't have any such desires at all, even to avoid leaving the house or rising from bed. If there is no level on which she desires for her situation to change then I would struggle to believe that she is really having any unpleasant experiences at that particular time. Unpleasant experiences that the subject is entirely unaware of, on all levels, do not seem to be plausible instances of unpleasantness.

Sufferers of depression need not be suffering constantly, and sometimes they will have more neutral or pleasant experiences. Suppose our depressed subject locks herself away in her room and boots up her favourite videogames. She gets lost in them; she feels soothed and after a while has no awareness of any kind of negative feeling. There are many things that we could say of her and her situation: it's bad for her in the long run, she still has the debilitating condition of depression, and she's unhealthy. This may all be true, but to describe the experience she's having

at that time as unpleasant would still be false. At that precise moment, although she may be ignoring some of her troubles, she is nothing less than content. The desire account, on which an unpleasant experience is an experience the agent intrinsically wants to avoid, is therefore compatible with these kinds of counter examples.

Old age – response one

In the case of the slow degeneration of old age, and the 'suffering by stealth' that ostensibly attends upon it, there are once again two possible ways that my response could go, depending on the details of the situation. Firstly, suppose that the subject's aches are noticeable. In this situation I can account for these aches by saying that there is always among the subject's everyday experiences some that she would desire to avoid, just as in the case of depression. She could be happy overall with a trip to the shops while harbouring a desire to avoid the unpleasant feelings in her knees as she does so, for example.

Old age – response two

On the second interpretation of the ageing example, the subject has no real awareness of anything negative going on at all, which seems to be the interpretation that Bramble gives.[110] This case seems, once again, to not be a plausible case of an unpleasant sensory experience. For the person these symptoms have crept up on, who is perfectly content with their lot, nothing unpleasant seems to be happening to them. Their aches are so subtle that the subject cannot even notice them, let alone find them unpleasant.

In the thought experiment that Bramble suggested, we imagined the subject being transported into their younger body and noticing how much easier, lighter and better it is. Just because a person can go on to experience life in a much better way, that doesn't mean that any part of their experience was necessarily unpleasant in the first place. To stick with the theme of improbable thought experiments, suppose a young and healthy person also develops the ability to fly. They'd be able to get around with even greater ease than they had before and their experiences might be generally far more pleasant. But however much better a situation might become, that doesn't mean that the experiences are necessarily unpleasant to begin with.

[110] Bramble (2013) p.206.

All cases of reflective blindness seem to fall down one of two sides of the trap: either they are not cases of sensory unpleasant experiences at the time; or the subjects are still able to have desires about the experiences without being aware of some other part, or way of characterising, what they're experiencing.

Depression – further thoughts

Depression is a particularly interesting example of an unpleasant experience, and it's one that deserves more discussion than just as a case of so-called 'reflective blindness'.[111] In fact, there's a second way in which it might be a counter-example to the desire account. Sensory unpleasant and painful experiences need to be accompanied by a desire for those experiences to stop, but a depressed state can be devoid of anything that might look like motivation[112]. Depression can be like an experience of having your motivations and your will drained, and characteristically leaves its sufferers unable to leave bed, to see friends, to eat, etc. This is the way that Rachels brings up depression as a counter-example: "[s]ome depressives have no impulse or only a slight impulse to change their condition, perhaps because they cannot imagine feeling happy."[113] In this section I'll discuss this version of the counter-example in more detail, and go on to show that the desire account I've been arguing for is not only not refuted by depression, but actually is supported by it and enhances our understanding of it at the same time.

Firstly I will mark a distinction between motivation and desire. Above I mentioned that it would be best to refer to the account as the desire account rather than following on from Heathwood and calling it the motivational account. This should be kept in mind when discussing counter-examples: to provide an example of subjects having no motivation isn't enough if they still have the relevant intrinsic *desire* for an experience to stop. Assuming that motivation and desire can come apart, then when they can it's the latter that should matter (and if they can't, then the example can be formulated in terms of 'desire' anyway, so this shouldn't make a difference). Times when they *might* come apart might be the kind of time when a counter-example could sneak in.

Suppose I'm in the office and I'm hungry in such a way that it constitutes a mildly unpleasant experience for me. The desire account accurately describes me as desiring for that experience – the hunger – to stop. But it's plausible that I might desire to not be experiencing

---

[111] For others who've tried to explain the link between depression and reasons / value, see for example Goldman, (2009) p.106.
[112] Ratcliffe, for example, refers to themes in descriptions of depression as including a "loss of hope" or experiencing the world as lacking "enticement" in Ratcliffe, (2015).
[113] Rachels, (2000) p.192.

hunger while also not being *motivated* to do *anything* about it. There are several reasons why this might be the case. The desire to stop being hungry might be in tension with several other desires I have, such as to work or to stay sat at my desk. The intrinsic desire to stop being hungry might not be my most vivid desire at the time, and it might not be motivating. Furthermore, it might be controversial that one can be motivated when there's no clear path to achieving the thing desired, and this might often be the case for unpleasant experiences. If I can't think of a way to satisfy my hunger then there won't be any specific actions that I would be motivated to take. Because of the way that desire and motivation might come apart it won't necessarily be accurate to use a lack of *motivations* in certain circumstances as a counter-example to the desire account.

There are times when depression is unpleasant (an understatement, maybe), but the subjects aren't *motivated* to do anything. To the extent that the experience is painful or unpleasant to the subject, though, she will still have the intrinsic desire for that particular experience to stop.

Having made these clarifications, I think there's still a way in which depression might be a counter-example. Rachels is right to describe depression as something which can sap someone of motivation, but he would *also* be right if he described it as something which can sap someone of desires. One of the common symptoms is a kind of emotional numbness which can affect the agent even regardless of factors that might normally make them happy. If one of the two main symptoms of depression is a depressed mood, then a reduced capacity to be happy is the second:

> To receive a diagnosis of Major Depressive Episode, a subject must suffer from five of […] nine symptoms in a two week period (*including* either or both of depressed mood or *diminished interest or pleasure in almost all activities*)[114]

To say that depression can reduce a subject's desires, even the intrinsic kind that I am interested in, is not a counter-example to the desire account but an example that *supports* it. The desire account helps us to explain a feeling of reduced desire in a depressed subject: it is a reason why they cannot find pleasure in activities they enjoy under normal circumstances, such as seeing friends, going outside, writing philosophy – because they have a reduced capacity for feeling desires.

The desire account provides a picture of depression that ties the changes in subjects' desires to the unpleasantness and lack of pleasure from experiences. Depression affects the ways

[114] Murphy, (2015) (emphasis my own). See also the American Psychiatric Association's DSM-5.

that an agent has desires in a way that skews them negatively. Subjects can be prone to feel desires more strongly if they're negative ones already, such as the kinds of desires that constitute unpleasant experiences: intrinsic desires for a certain and current experience to stop. This is when depression is the most painful to the subject. The subjects can also be prone to feel desires more weakly if they're positive desires; an intrinsic desire for an experience to continue may be felt weaker than it would under different circumstances – if felt at all – so the subject feels less pleasure, neutral, or even finds it unpleasant in circumstances they might once have enjoyed.

I should note that this isn't an attempt to diagnose depression, and I have nothing to say about its physical causes. What I have argued is that the desire account is far from disproven by it, and that it rather actually provides us with a plausible way to understand and think about what we already know about depression.

2.2.5 Other objections

This section will turn to a different kind of counter-example: ones in which the subjects have the supposedly relevant kinds of desire but the experiences that they're having aren't the kind that should be described as pleasant. After all, the kind of claim that the desire account is making must go two ways: if a certain kind of desire is what makes an experience pleasurable then those kinds of desires must always be pleasurable.

Sidgwick's own view of pleasure was unclear, in that he has been interpreted to have a variety of different views.[115] Whichever of these is true, some of the examples he gives in discussing the matter could be seen as counter-examples for the desire account. Hurka, for one, thinks that Sidgwick's examples refute the desire account, and he also adds one of his own to finish the job. I'll turn to these two next.

Blueness and exercise

In trying to pin down various views on the nature of pleasure, Hurka gives the following example,

[115] Shaver catalogues a variety of interpretations in Shaver, (2016) and Hurka also describes it as 'unstable' in Hurka, (2014) p.195.

> […] imagine that blue is someone's favourite colour, so he wants to have sensations of blue. On the [desire account] these sensations are pleasures, and they are so even if he has no feelings whatever about them, his mind containing just an awareness of blue and a desire that it continue. This does not seem right; though a sensation of blue may cause pleasure in someone who likes blue, it is itself just a sensation of blue.[116]

The desire theorist needs to either justify why this particular example is not an example of a pleasurable experience, or concede that thinking about blueness in this way is pleasurable.

Consider the kind of experience of blueness that would be needed under the desire account for a pleasurable experience. The subject would have to be experiencing blueness, let's say by looking at the walls in a blue room. The subject would have to have an intrinsic desire, while experiencing the blueness, for that particular part of the experience to continue. This is certainly not 'just' a sensation of blue, as 'just' a sensation of blue wouldn't come with any intrinsic desires. I have sensations of blue fairly often, and do so without desiring for those experiences to continue, for one because blue is not my favourite colour. Since it's not 'just' an experience of blueness, it seems like the most sensible thing to add is that it a *pleasant* experience of blueness.

Furthermore, I'm not sure what Hurka really means by saying that one can have pleasant sensations without having any *feelings* about those sensations. It sounds a bit like Hurka might be trying to distinguish between the bare physical sensations and the pleasure that those sensations give, as I've discussed above in terms of distinguishing the different possible meanings of pain, and as Parfit discussed in separating hedonic likings and meta-hedonic desires from the sensations on their own. If this is the case, then this still doesn't provide a plausible counter-example to the desire account. The account argues that all pleasurable sensations are desires, and makes no claims about the bare physical sensations themselves, to the extent that they could be separated.

Out-of-proportion responses

Returning to Sidgwick, the most memorable example that he contributes to the debate is the example of tickling. He notes that pleasures are not greater or lesser "exactly in proportion as they stimulate the will to actions tending to sustain them."[117] Sometimes there is pleasure or pain but no stimulus to act, either because one has the experience one wants or one becomes accustomed

[116] Hurka (2014) p.195.
[117] Sidgwick, (1907) p.126.

to the pain.[118] But such stimulus to act is not what the desire account is about, as clarified above. The thesis refers not to actual motivations towards actions at all, but rather to intrinsic desires for the continuing or stopping of an experience.

Summary

What it is to find an experience pleasurable is to intrinsically desire, at the time of the experience, for that desire to continue. This is the desire account of pleasure. I've demonstrated, as best I can, that all examples of the relevant desires and all examples of pleasant and unpleasant experiences match up appropriately to the account.

2.2.6 Returning to value

The main job of this section has been to argue that pleasurable and painful experiences are subjective. They are subjective because what it is for an experience to be pleasurable or painful is for the subject to have a certain kind of intrinsic desire; if that subject doesn't have the desire, the experience is not pleasurable or painful. This means that Parfit cannot argue against subjectivism on the grounds of pleasure and pain being separable from any subjects' desires.

Sobel agrees that to tie-up hedonic likings and meta-hedonic desires this way would solve the problem for the subjectivist.[119] He notes that meta-hedonic desires would be exactly the kind of authoritative state that could 'confer value' in a way that Parfit argued they couldn't,

> If likings were a kind of desire, subjectivists could account in a natural way for the reason-giving power of such states in a way that fits well with their broader approach. In other words, if likings were desires, they would be just the sort of desires that subjectivists can most plausibly grant authority to; namely those desires which are accurately informed about their object.[120]

Parfit argued that hedonic likings and meta-hedonic desires were two separate parts of a mental state, and that proponents of subjective value were getting the two confused. Meta-hedonic desires

[118] Shaver describes Sidgwick's opinions in this way in Shaver, (2016) p.901.
[119] Lang also hints at this approach as a response to Parfit, saying "the hedonic disliking is not innocent of association with desire." Lang, (2012) p.303.
[120] Sobel, (2011) p.59-60.

were desires, but could not confer any value. Hedonic likings were what conferred value, but they were *not* desires. In this section I have argued against the distinction by arguing in favour of the desire account of pleasurable and painful experiences: an account on which what it is for an experience to be pleasurable or painful (or for an agent to have a hedonic liking or disliking) is the same as for an agent to have a certain kind of intrinsic desire for that experience to continue. Now I'll tie up a few loose ends with my response to Parfit's argument.

I've argued that there's no real distinction between hedonic likings and meta-hedonic desires, but I've not explicitly addressed two reasons that Parfit gives for this distinction, so I'll do that now.

Likings vs desires: desire-satisfaction

Parfit's first concern is to do with desire-satisfaction. In describing the difference between the two, he says,

> What we dislike is some sensation. What we want is not to be having a sensation that we dislike. Our desire could be fulfilled either by our ceasing to have this sensation, or by our continuing to have it but ceasing to dislike it. No such claims apply to dislikes, which, unlike desires, cannot be fulfilled or unfulfilled.[121]

I've spoken a bit already about the difference between the desire account and accounts that explain pleasurableness in terms of desire satisfaction. I made the case that the nature of the specific desire in question – that the desire is intrinsic and directed at a current experience, for that experience to either stop or continue – is such that the satisfaction of the desire is not what's important. It's certainly not important in terms of whether it constitutes the pleasure or painfulness of an experience (since it's the desire – not its satisfaction – which constitutes the pleasure or painfulness).

It's not clear why the ability of a desire to be fulfilled should mean that that desire is a completely separate entity from an experience of pleasure or pain. It doesn't seem so outrageous to say that there's a sense in which pleasure and pain experiences can be fulfilled or denied: for pleasure to continue would be fulfilling, for pain to stop would be fulfilling. As Thomson tells us,

[121] Parfit, (2011a) p.54.

some mental states have 'correctness conditions'.[122] Pleasurable and painful experiences on any account that separates them from the mere sensations (and so Parfit's included)[123] come with some kind of way to analyse why it is that pleasure is good and pain is bad, and having desires that go in opposite directions (desires to continue versus desires to stop) seems like a good way to account for that.

In fact, Aydede states that the ability to explain how pleasure and pain can be so opposing is a necessary requirement for a theory of the two, and opposing desires is a way to do so.[124] He calls this the puzzle of 'Opposite Valences', and says that the use of desires, lack of desires, and desires for the cessation of experiences means the puzzle is "easily solved".[125]

But even if this weren't persuasive, it seems like a very small bullet to have to bite. If it doesn't sound right to say that a dislike cannot be fulfilled or unfulfilled, but it does sound right to say that desires can be fulfilled or unfulfilled, this doesn't seem like a good argument for why dislikes can't be constituted by desires.

Likings vs desires: future directedness

Parfit's second reason to distinguish between hedonic likings and meta-hedonic desires is to do with whether they can be directed towards different times,

> Unlike our meta-hedonic desires, our hedonic likings or dislikings cannot be aimed at the future, or at what is merely possible. That is another reason why I do not call these mental states desires.[126]

This problem is easier to solve. The desire account is already very specific about the kind of desire that's under discussion. Meta-hedonic desires, the ones that constitute the pleasure or painfulness of an experience, do not encompass a wide range of desires, but only desires that are for an already occurring experience to continue or stop.

---

[122] Thomson, (2008) p.116.
[123] I discussed how Parfit separates hedonic likings/dislikings from the bare sensations in **section 2.1**.
[124] Aydede, (forthcoming a) p.11.
[125] Aydede, (forthcoming a) p.12-13. This isn't necessarily a point in favour of the desire view over Parfit's own view, though, as Parfit agrees that in most cases pleasure and pain are accompanied by meta-hedonic desires anyway.
[126] Parfit, (2011a) p.54.

Of course, the chapter as a whole is about finding out whether *what's valuable* is subjective, not just about whether *pleasure and pain* are. But this section has still been an important step in that argument: it has both refuted possible objections that require that pleasures and pains are valuable, and it's laid the foundations for an argument about whether pleasures and pains are valuable.

2.3 Future Desires

The second argument of Parfit's that I put under the microscope is his 'agony argument'. According to this argument there must be more to what's valuable than any subject's current set of desires because there are certain things which may definitely be valuable or disvaluable but are in no way related to that subject's current set of desires. He gives the example of future agony, and argues that any thesis which doesn't treat future agony as a source of disvalue is an implausible account. In this section I will address this argument with help from Street's work in her paper 'In Defence of Future Tuesday Indifference'. I will agree that future agony does pose an interesting problem for the work I've done so far, but conclude that in the very rare cases when Parfit's examples would be applicable, we should bite the bullet and accept that future agony has no effect on either a subject's current reasons or what they value.

Future agony poses a particularly interesting problem for my account of value because, as I discussed in **2.2**, agony is something that the subject will *necessarily* desire to stop at the time that it's occurring. The nature of pain is such that there can be no doubt that if someone is going to experience agony that it will be, at the time of the experience, something they would want to stop. But nowhere on the subjectivist account that I've been arguing for is there a necessary connection between current desires and *future* agony. It's not conceptually possible to be in agony and not have an equally strong desire for the agony to stop, but it *is* conceptually possible to know you'll be in agony in the future but to not have any kind of desire to avoid it.

Sharon Street takes an in-depth look at Parfit's criticism and argues that after careful consideration the kinds of situation in which future-Tuesday indifference would occur make the intuitions Parfit relies on go away.[127] She argues that meta-ethicists such as Parfit (as well as Hume with his man who would prefer the destruction of the world to a prick on his finger, Rawls and his person who loves counting blades of grass and Gibbard with his anorexic)[128] should not use

[127] Street, (2009).
[128] Street, (2009) p.273.

such insufficiently developed examples to make important meta-ethical points. The characters in question are actually so far removed from what's familiar that the examples are likely to be unfairly pumping the wrong intuitions.

Looking specifically at Parfit's agony argument case, Street puts forward two possible ways that the character with future-Tuesday-indifference could be: either the character's indifference continues throughout the Tuesday or the indifference suddenly goes away at some point leading up to a painful event, perhaps as it turns to 12am on Tuesday or perhaps just as the pain begins.

In the former case, Street argues that there are two lessons to take away from examining the case in close detail, the most important of which I believe is this:

> Because of the nature of pain, examples involving pain require especially careful consideration. Pain is not just some ordinary object of our evaluative attitudes, but rather a phenomenon which by its very nature seems to involve evaluative attitudes directed at certain bodily sensations. This complicates attempts to see whether an attitude-dependent or attitude-independent conception provides the best account of the reason-giving status of pain.[129]

Indeed, as I said above the nature of pain is what makes Parfit's agony argument such an interesting case. I've already argued in **2.2.2** that if someone has absolutely no intrinsic desire for an experience to stop then it isn't a painful experience, so the possibility that the subject would be indifferent even up to and during the point of agony is ruled out. This could only be agony in a sense unlike anything we think of as agony; it would be an agony that just doesn't matter. This version of Parfit's counter-example is therefore unpersuasive.

The most interesting case left, then, is one where the subject will care very strongly about the agony when it's happening, but simply has no desires about it at a certain point in time beforehand. Parfit claims that this future agony should still be reason-giving, still be disvaluable, but the subjectivist cannot account for why that would be. This, I think, is how to best understand Parfit's agony argument, because it poses the greatest challenge for the subjectivist.

My response consists firstly in exploring whether the special nature of painful experiences can save the subjectivist. I will then conclude that they can't completely save the subjectivist from

[129] Street, (2009) p.288.

having to bite the bullet, but they can do so almost well-enough, and in a way that is actually pretty painless.

As I mentioned above, pain is a particularly interesting case because of the conceptual connection between pain and desire. If there will ever be a conceptual connection between what's valuable, and what we have reason to do, with our *future* desires and circumstances instead of our present ones, then painfulness seems to be where we'd find that connection. After all, if you're certain that you will be in pain then you can be certain that you'll desire for it to stop, when you're experiencing it. It's something that we can be certain will be disvaluable at the time of experiencing it. All we would need is to be able to connect that future desire, somehow, with currently finding it disvaluable.

For the subjectivist account I've been defending, the necessary connection needs to be between a subject's actual, current set of desires and what's valuable to her. Otherwise, the valuable thing is not valuable *to her*, currently, but rather valuable to some future version of her. The only way that the subjectivist would be able to account for a necessary connection between future desires and current value, then, would be to connect those future desires with her current desires.

There's a case to be made for subjects having desires to avoid future pains. This is clear when we think about the ways that we could describe that kind of desire. Such desires, for example, might look like a desire to avoid pain (at any time), or a desire to avoid things at any time that you'll (at the time) desire to avoid. Although they sound a bit unwieldy, those are both desires that I am happy to describe myself as having, for example.

But there just isn't a way to argue for the conceptual necessity of these desires in the same way that I argued for the conceptual necessity of desiring to avoid current pains. After all, on my account, it isn't even necessary that subjects always desire to currently avoid pain. They only have this desire *if they're in pain* at the time, and when they are the desire is only directed towards that specific pain that they're currently experiencing. It looks like there's no way, on the subjective account, to necessarily guarantee that subjects will have any desire to avoid future pain, and no way to necessarily be able to describe that future pain as disvaluable to the subject.

If the subjectivist wants to be able to describe future pain as disvaluable to a subject, then, it will need to rely on contingent facts about what the subject does happen to desire. And that'll be what makes up the rest of my response to Parfit's future desire argument. I'll make two claims: firstly, that in nearly every case subjects *will* desire to avoid future pain; and secondly, that a subject

who doesn't have such a desire, no matter how broadly I take desires to be, is such an unfamiliar creature that we shouldn't be put off by subjectivism giving us this apparently unintuitive result.

Most subjects will have at least some desire to avoid agony in the future as well as the present. Once again, this becomes particularly clear when we remember that desires here are taken to be very broad, and inclusive of desires that feature in the background of our thoughts as well as those we feel most vividly at any given time. Furthermore, the case becomes clearer when we think about what might be otherwise problematic cases. In **section 2.1** I mentioned other examples of when subjects behave imprudently, in a way that might seem analogous to Parfit's 'Future Tuesday'-type cases: when a subject finds it far easier to prioritise their immediate happiness, for example, over their long-term health. But these kinds of cases are *not* analogous with Future-Tuesday cases at all. Rather, these are cases where the subjects *do* have the relevant desires (such as to be healthy), but the desire just doesn't strike them as hard, at that particular moment, as the desires for more immediate gains. When a subject finds themselves unable to put down the pizza menu, we wouldn't say that she doesn't have any desire to be healthier and better at saving money, for example, but either that she's weighed up her desires and still found the pizza to win out, or that she's being weak-willed in some way.

The kinds of subjects in Parfit's cases are very different creatures. Street refers to them as ICEs, or "Ideally Coherent Eccentrics".[130] She argues that once we've ruled out subjects who do have the relevant desires (and are perhaps just failing to act on them or express them), then what we're left with is a very strange and unfamiliar subject. Such a subject would have to have peculiar battles with their future selves over what to do, in a way that almost treats their future self as a different person altogether.[131]

The idea that subjects might exist without such desires is incredibly unlikely to ever be relevant. If there ever are any subjects who don't desire to avoid certain future agonies, then I am happy to bite the bullet and agree that those subjects just don't have reason to avoid that agony until such a time when they do have such desires.

As a final note for this section, this rarity of subjects who don't desire to avoid their own agony is also likely to be the case when we think about subjects and their desires towards morality. Reasons internalism cannot impose normative reasons to avoid agony on any agents unless those agents already have desires to avoid future agony, nor can it impose normative reasons to be good

[130] Street, (2009) p273. Describing them as "ideally coherent" allows us to talk about all of their desires, including those featuring in the background.
[131] Street, (2009) pp.281-292.

on any agents unless those agents have a desire to be good. As I argued in Chapter 1, this is a necessary sacrifice for a plausible theory of normative reasons. As it happens, cases where those desires are missing are fortunately incredibly rare.

Conclusion

This chapter began by explaining the relationship between reasons internalism and value subjectivism. Reasons internalism is the thesis that an agent's normative reasons for action are contingent on her current set of desires, and value subjectivism, that what's valuable to a subject has that same kind of desire-based connection. This chapter defended both of those theses against two significant objections from Parfit.

The first objection was that the subjective nature of valuable experiences such as pleasure and pain could be separated from what made them pleasurable or painful. I argued that they could not because what it is for an experience to be pleasurable or painful *is* for the subject involved to have a certain kind of desire. This took up the bulk of the chapter, and I fended off what I came to call the 'desire account' of pleasure and pain against a range of possible counter-examples, including those of 'reflective blindness' and 'depression'.

The second objection was based on the possibility that our desires can come apart from what's best for us, and that there are some extreme cases of this that seem too implausible for the subjectivist. I responded by showing that these cases were too rare to be a problem, and that cases when subjects really do have such unusual sets of desires (such as someone who has no desire to avoid pain that will occur on a future Tuesday) are so unusual that it's not surprising that such subjects will have equally strange reasons to act and find equally strange things valuable.

Chapter 3.

What We Ought To Do:
Against Categorical Imperatives

Introduction

Chapter 3 will be the home of my first arguments for the view that all normative 'oughts' are subjective. I'll describe them as 'hypothetical imperatives' for most of this chapter, to match the language of Kant and Foot. I'll argue that they're not as different from 'categorical imperatives' as it might seem at first, and that the only thing categorical imperatives have that they don't is the ability to apply to agents regardless of their desires. This, I'll argue, isn't a quality that a normative imperative should have.

Wider context

The first half of my thesis aimed to explain our normative reasons for action. In it I argued that any agent's normative reasons are necessarily related to their desires, broadly construed. The second half of my thesis will build on this work. I will argue that it's not just our *reasons* that are contingent on our desires but also what we *ought to do*. That is, all of what an agent ought to do is necessarily dependent on what they desire. I'll discuss two main competing kinds of ought; the latter half of Chapter 3 will deal with categorical imperatives and Chapter 4 will deal with 'overall' oughts: oughts that aren't dependent on any single desire or set of desires but are rather a product of an overall judgment.

This chapter

This chapter is split into three main sections. The first, **section 3.1**, will explain what I mean by hypothetical imperatives. I'll explain their structure, justify why I've chosen to define them as I have, and discuss and reject two similar arguments against a theory that understands oughts as hypothetical in this way: the 'too-many-reasons' objection and the 'bootstrapping' objection. The former, I'll show, has already been responded to well by Schroeder, but I will go further than him and demonstrate that many of the counter-examples that he thinks we should accept aren't counter-examples after all. For the bootstrapping objections I'll give a new response, which demonstrates why we shouldn't worry about the structure of hypothetical imperatives 'bootstrapping' normativity into existence. In doing so, I'll be able to give a good explanation for why it is that the normative concepts I've been discussing in this thesis are actually normative, where it is they get their force from: because of the relation to the agent's desires.

Section 3.2 will introduce the more complicated concept of categorical imperatives. I'll go through five different criteria that are said to distinguish categorical imperatives from merely hypothetical ones: importance and dignity; applying in virtue of the agent's rationality; requiring us to perform actions for their own sake; applying with authority and inescapability; and applying to us categorically. I'll show that hypothetical imperatives can meet the first four criteria and that the fifth isn't plausibly a criterion for normative 'oughts' after all.

Some philosophers, such as Joyce and Mackie, have worried that the fact that moral imperatives cannot apply to us both normatively and regardless of our desires means that we should be error theorists about morality. In **section 3.3** I argue that such categoricity isn't an important part of our moral discourse after all, and that my account of normativity as subjective is compatible with moral realism.

Reasons and oughts

Before I begin my arguments, I will say something about the relationship between reasons and oughts. They're both normative concepts. They're more than just concepts that explain just why things should happen physically or mechanically, ones that (for example) explain how the books are going to fall over because the cat is about to knock them, or that Alma is going to cancel her

plans because she's depressed. Instead they try to capture something about why agents *should* choose certain options over others: why they should believe certain things, cultivate certain virtues or act in certain ways.

According to convention, a reason weighs in favour of or against something,[132] and I take it that what an agent *ought* to do is a kind of conclusion of their reasons. When a certain set of reasons is weighed up, the action with the most weighted reasons (unless there are other defeating conditions) will point towards the action that the agent ought to do.

There are many different kinds of reason. This is the case even when we've ruled out all of the different kinds of labels - explanatory reasons, motivating reasons, objective reasons, etc.- and settled on 'practical normative reasons' (see Chapter 1 for more on this). Even then, there are different kinds of reason that it can be helpful to distinguish between. For example, there are moral reasons, prudential reasons, legal reasons, and social reasons. Reasons of friendship, reasons of science, reasons of being a good bearded dragon owner, reasons of faith. The list is long, overlapping and no less arbitrary than our labels are of those concepts generally. Our legal reasons and our prudential reasons might often be the same, so might our moral reasons and our reasons of friendship, etc. Other times it might be helpful to talk about how our moral reasons might conflict with another kind of reason, and try to work out an appropriate action to take given their different weight and value.

The difference between each kind of reason is what desire (or set of desires) the reasons are dependent on. The moral reasons are the reasons we have because of our moral desires, whatever those turn out to be. An agent's prudential reasons are those which she has in virtue of her desires to be prudent (or, at least, her desires that align with prudence), her bearded-dragon-owner reasons are those she has in virtue of her desires to be a good owner of her bearded dragon(s) and those related desires, etc.

We can make these same kinds of distinctions between the 'conclusions' of reasons: the things that we ought to do. Suppose that the conclusion of all of the moral reasons that I have at the moment is that I ought to attend a protest this afternoon in the centre of my city. Given the parameters of a finite set of options that I'm considering, and a finite set of reasons that I have in virtue of my desire to be a morally good person (if, indeed, you think that's the relevant desire for moral reasons), then that's what I ought to do: attend the protest. But that 'ought' is qualified, because it's the conclusion of just a specific set of my reasons: my moral reasons. It's what I morally

[132] I say 'convention', but Scanlon (1998) should also definitely take some credit.

ought to do, but the question remains open as to whether there are other things that I ought to do too. As a bearded dragon owner (and, implicitly, as someone who desires to be a good token of that type) then I ought to go to the reptile store and get a new UV bulb for my bearded dragon's vivarium. As a good friend I ought to stop by Bob's house and bring her a surprise afternoon burrito after she's had a difficult day. Each set of reasons can generate its own kind of 'ought'. And so, there are many kinds: moral oughts, bearded-dragon-owner oughts, and oughts of friendship, for example.

All of this will be explained and justified in more detail in the coming two chapters. For now, I just wanted to establish the kind of relationship I have in mind between reasons and oughts. Both, I will argue, are conditional on the desires of the agent in question. Both can come in a variety of different types, which ultimately depend on the desire(s) in question. The difference, then, is that what we ought to do can be a conclusion of multiple different reasons. It's a way of saying what we have the *most* or *strongest* reason to do, given a certain desire (or set of desires). Taking into account all of the reasons of friendship (reasons to decorate Sophie's desk for her while she's away, reasons to go online and make some new friends, reasons to bring Bob a burrito) the strongest reason, and the conclusion, might be that I ought$_{\text{(friendship)}}$ to bring Bob a burrito.

As a final clarification before I get on with the chapter, I want to mention the relationship between oughts and obligations on my account. The term 'obligation' has some stronger connotations than the ought concept that I want to understand. I want to understand oughts fairly broadly, and, will become clear in **3.1.3** there are things that we ought to do in some sense that we definitely ought not to do in some other (and perhaps more important) sense. For that reason, I'll avoid talk of 'obligations' for now (I'll briefly return to them in Chapter 4).

It might seem, given what I have just said about the connection between reasons and desires, like a picture of oughts which connects them necessarily to an agent's desires follows on quite easily from a picture which connects an agent's reasons in the same way. But such a picture of oughts comes across its own set of objections, and is worth arguing for. That's what the rest of this thesis will do.

3.1 Hypothetical Imperatives

Introduction

This section will talk about what it is for an imperative – that is, an 'ought' statement[133] – to be hypothetical. There will be three sub-sections, and the first of these (**3.1.1**) will begin by defining what I mean by 'hypothetical imperatives', what concept it is I'm hoping to pinpoint. **Section 3.1.2** will defend my own definition, and justify, firstly, why I exclude two particular qualities that other definitions might take into account, and secondly why my own definition is philosophically important and worthy of pursuit (particularly bearing in mind the topic of my thesis overall). Finally in **3.1.3** I'll discuss the problems of 'too-many-reasons'[134] and of 'boot-strapping',[135] both of which are worries about the implications of the structure of hypothetical imperatives as being implausibly obligating. I'll demonstrate why these worries don't affect hypothetical imperatives and, in the process, make the structure of hypothetical imperatives (specifically of their 'normative force') even clearer.

3.1.1 Defining hypothetical imperatives

Kant is responsible for a lot of the discussion on hypothetical and categorical imperatives, as the distinction played an important role in his account of morality. He describes hypothetical imperatives here:

> Hypothetical imperatives declare a possible action to be practically necessary as a means to the attainment of something else that one wills (or that one may will).[136]

And then a little later:

> Every practical law represents a possible action as good and therefore as necessary for a subject whose actions are determined by reason. Hence all imperatives are formulae for determining an action which is necessary in accordance with the principle of a will in some sense good. If the action would be good solely as a means *to something else,* the imperative is *hypothetical*; if the action is represented as a good *in*

[133] "All imperatives are expressed by an *'ought' (Sollen).* By this they mark the relation of an objective law of reason to a will which is not necessarily determined by this law in virtue of its subjective constitution (the relation of necessitation)." Kant, (2012) p.77.
[134] Also discussed by Schroeder in Schroeder, M. (2004).
[135] Discussion of this can be found in, for example, Finlay, (2014) and Kiesewetter, (2017).
[136] Kant, (2012) p.78.

> *itself* and therefore as necessary, in virtue of its principle, for a will which of itself accords with reason, then the imperative is *categorical.*[137]

Although some of my discussion in this chapter might not be true to his original meaning (more of that in the next subsection) he picked out a distinction that I find very useful in understanding what we ought to do. That is, there's a difference between two ways of understanding normativity: understanding normativity in relation to desires, and understanding it not in relation to desires.[138] This chapter (and my thesis as a whole) should make it clear how important that distinction still is in moral philosophy.

When there is an end that an agent could will, a hypothetical imperative (as I will understand it) describes what that particular agent ought to do to bring about that end. The hypothetical imperatives are, as I will understand them, very much the kinds of thing that do just what they say on the tin. That is, they're (1) imperatives (statements that purport to explain or command what to do) that are (2) hypothetical (conditional on something). The thing that they're conditional on is the desires of the agent, and the action is something that might bring about the desired outcome.[139] This is simply how I will define them. To make it particularly clear, they can take this form:

If A desires X, then A ought to φ

Where A is an agent, X the state of affairs that they desire to come about and φ is an act that might bring about that state of affairs.

---

[137] Kant, (2012) p.78.

[138] There could also be hypothetical imperatives that are contingent on something other than an agent's desires, ones contingent on states of affairs obtaining, for example. "If it's past 9pm then you ought to go to bed" might be an example of this. But for the purposes of my thesis I'll take 'hypothetical imperatives' to refer to the kind contingent on desires.

[139] By saying 'might' here I mean to show that the exact relationship is complicated, just as the analogous relationship was in Chapter 1 when I discussed the link between reasons to act and the states that agents desired. It could be that hypothetical imperatives apply when the action *will*, as a matter of fact, bring about the desired state of affairs, or it could be that they apply when the agent in question *thinks* that the action will bring about the state of affairs, or it could be some middle-ground between the two. My own position, of course, is the latter, given that hypothetical imperatives are, I believe, a conclusion of certain sets of reasons, and those reasons are sometimes informed by an agent's beliefs and sometimes informed by what's true. I won't repeat the arguments here.

Foot also plays a significant role in the debate about the two kinds of imperative,[140] and in order to demonstrate what a hypothetical imperative might look like in practice I'll borrow some examples from her:

> Sometimes what a man should do depends on his passing inclinations, as when he wants his coffee hot and should warm the jug. Sometimes it depends on some long-term project, when the feelings and inclinations of the moment are irrelevant. If one wants to be a respectable philosopher one should get up in the mornings and do some work, though just at that moment when one should do it the thought of being a respectable philosopher leaves one cold.[141]

Here we have two hypothetical imperatives. In the form I specified above, they look like this:

> If A desires hot coffee, then A ought to warm the jug.

> If A wants to be a respectable philosopher, then A ought to get up in the mornings.

These examples also serve to demonstrate the point that the desires in question don't need to be the most strong and obvious desire at any given time in order to still play a role in hypothetical imperatives. When I wake up in the morning and all I can think about is my desire to stay in bed, that doesn't mean that this is my only desire. I still have projects I desire to continue, people I care about and desire to do well, etc., and so staying in bed isn't the only thing I would have a reason to do nor the only thing I ought to do.

Conditional desires as not-explicit

It's worth saying that hypothetical imperatives will be far more commonly used than you might immediately think from their description. This is because the conditional part of the imperative will nearly always be implicit, which explains why we don't hear people explicitly referencing the

---

[140] Her paper 'Morality as a System of Hypothetical Imperatives' (1972) is a large influence on my arguments in **3.2**, which will become clear. She did later change her views on the matter in Foot (2001). I've not yet changed mine.

[141] Foot, (1972) p.306.

conditional part of the imperative when they talk about, or use, them. Finlay makes this point in his book *Confusion of Tongues*[142] and gives a variety of examples of why the conditional desire (or the 'end' which is desired) might not be explicitly stated. Here I'll cover some of these.

One reason why it won't be common to explicitly mention the relevant desire or end is because they are obscure or hard to pin down. Finlay says,

> [Ends] can be obscure due to a variety of factors: multiple ends might be equally salient, the conversational end might only be vaguely recognized, or charitable interpretation may rule it out, and so on.[143]

Sometimes there might be multiple desires that the imperative follows from, as in the case of 'you should get up in the mornings'. The speaker may well assume that the subject has several ends that would be satisfied by getting up in the mornings, and can refer to these without explicitly mentioning them or even knowing exactly what those ends are. After all, it's not just my I desire to be a philosopher that should get me out of bed, but a variety of other desires, projects, and ambitions. So, we have at least two reasons already why an agent might avoid being explicit about the desires and/or ends: firstly, is that given multiple ends it may not be clear which is most relevant, or important, at any particular time, and so too much effort to determine which to state. Secondly because the sheer number of ends often makes it more trouble than it's worth to list them. Thirdly, Finlay asserts, the reason to obscure the relevant end is that it may be particularly complicated or hard to explain (such as with aesthetic ends).[144]

There are also ends that speakers and listeners alike will take for granted. We both know that we both want some coffee, so there's no point in you explicitly saying that when you tell me to go and warm the coffee jug. Explicitly mentioning the ends in many of these cases would be redundant.

---

[142] Finlay, (2014) p.146-175. He describes the examples of implicit hypothetical imperatives in much more detail than I do. Davidson also makes a briefer but similar point in Davidson, (1963) p.688-689, saying for example that "If I say I am pulling weeds because I want a beautiful lawn, it would be fatuitous to eke out the account with 'And so I see something desirable in any action that does, or has a good chance of, making the lawn beautiful.'" Not all of the steps that go between actions or imperatives and the relevant desires will be explicit in most cases.

[143] Finlay, (2014) p.146.

[144] Finlay, (2014) p.150.

Another reason why an end might be implicit is because obscuring the relevant desire might be *more likely* to persuade the subject of the right course of action.[145] Sometimes it may be beneficial to leave the ends obscured because you might be more likely to persuade someone they have a reason to do something if you don't list all of the possible reasons they might have, the possible ends they have that might be fulfilled by them doing that thing. This might mean they're less able to dismiss these ends one-by-one, for example. Leaving the ends implicit might imply that there are somehow *more* ends that would be achieved, more desires that would be fulfilled, than there actually would be. In other cases to leave ends implicit could perhaps even sound threatening ("just do it!").

I'll leave this discussion with one more relevant quote on the matter from Finlay:

> While no particular ends may be uniquely salient, in these cases there are still salient persons, and thereby salient sets of ends, being those desired or intended by the speaker and/or audience.[146]

This quote does a good job of putting this discussion in the context not just of this chapter but of one of the most important arguments in my thesis as a whole. Ends are relativized to desires because they're relativized to agents, and those agents are not the kind of things we can separate from their desires. I'll have more to say on this in **3.2**, and I have already made the case in some detail in Chapter 1. For now, this gives us a further reason to see why the lack of explicit and singularly salient ends is not a problem for an account of 'oughts' as hypothetical imperatives.

I've now given my definition of hypothetical imperatives and, with help from Finlay, given several reasons to think that hypothetical imperatives are commonly used, despite the ends or desires not being explicitly listed in a lot of cases. I'll now say a bit more on the 'conditional' part of the imperative.

The normative 'force'

Hypothetical imperatives, to their credit, are very transparent when it comes to seeing what normative force they have. It's easy to see why the imperative applies to the agent and why it might be that the agent has a reason to perform the action. On my account the 'force' that obliges, requires or commands the agent to act comes from the agent's own desires. If the desire isn't

[145] Finlay, (2014) p.149-150.
[146] Finlay, (2014) p.146.

present (perhaps because it's subsided or it was never there to begin with) then the imperative doesn't apply to that agent. That's not to say that the agent has no reason to carry out the action, of course, because they might have *other* reasons to do so, there might be other hypothetical imperatives that apply to them. It just means that that particular hypothetical imperative doesn't apply to them, and so doesn't give them a reason, at that time.[147]

Using these examples we can see better how the normative force behind the hypothetical imperatives can come and go.[148] When I want hot coffee then the imperative 'if you want hot coffee, then you ought to warm the jug' applies to me. I should, indeed, warm the jug, and I should do so *because* I want the coffee to be hot and this is a way to make that the case. As soon as I stop wanting hot coffee then my reason to warm the jug just goes away; I no longer ought to do it unless there are other imperatives, other reasons or ought statements that apply to me. This can be seen most obviously in the simple cases like the coffee case, but can be generalised to any hypothetical imperatives.

The only reason why it's so much more difficult to escape the normative force of a more important ought-statement (like the one that tells me to get out of bed in the morning and do some philosophy) is because it's much more difficult in these situations to escape that kind of desire. The desire to become a philosopher is a much stronger, more long-term, long-standing desire that doesn't go away even if I can't feel it when I wake up, even if it's not the first thing in my thoughts in the morning or the thing that my attention is centred on at that exact time. This is often the case for our moral desires.[149]

I've now covered the basic definition of a hypothetical imperative. It's an ought-statement, one that's conditional on the agent having a certain kind of desire. It most obviously takes the form of "If A wants X, then A ought to φ", but in everyday speech the antecedent is usually implicit. We looked at a couple of examples, and saw that the desires that the imperatives are conditional on can be fleeting or more long-standing. In either case, the ought-statement only

---

147 I talk about hypothetical imperatives providing reasons at various points here, but it should be noted that what I really mean is something more like the fact that the hypothetical imperative describes, points out or states a reason that the agent has. The existence of a hypothetical imperative doesn't give the agent any reasons to act that they didn't have already. Talk of hypothetical imperatives giving reasons is just a natural-sounding way to talk about the reason the hypothetical imperative indicates.

148 When I talk about normative 'force', what I mean is the kind of thing which makes the difference between, say, something one could do and something one *should* do. It can come in degrees, and there are cases when an agent really *really* ought to do something (perhaps, for example, they have a significant obligation) and other cases when the normative force is quite weak; when they should do something but it doesn't matter so much, and there are other alternative actions that would be almost as good.

149 I'll have more to say on normative force in the subsection on bootstrapping in **3.1.3**, and more to say about the escapability of certain desires and hypothetical imperatives (particularly moral ones) in **3.2.4**.

applies when the condition is met. The condition can be dropped easily when the desires are more fleeting, but this is less true of stronger and more resilient desires. In the next two subsections I will make some more clarifications about hypothetical imperatives. The first of these will address two potential competing definitions to the one I've given.

3.1.2 Alternative ways to define hypothetical imperatives

There might be at least two problems with the way I've defined hypothetical imperatives. Firstly there may be exegetical problems: problems to do with what Kant originally meant and whether my definition is true to what he meant. Secondly, there may be competing definitions of hypothetical imperatives (ones that are perhaps closer to what Kant meant) and I should defend why the one I've given is the best one to use given my overall project. Here I'll address both of these concerns. I'll first say something about why I won't attempt to figure out Kant's intentions, and then I'll discuss two possible features of a definition of hypothetical imperatives that I have consciously chosen not to include in my own definition.

Although I began with a quote from Kant, perhaps the definition that I've introduced above is not completely similar to the one that he had in mind, and perhaps other passages in Kant bring that out. But this is not a thesis on Kant, and for most of this chapter I want to concentrate only on hypothetical imperatives as I've defined them. Beyond brief discussion in this section I will remain fairly silent on which is the most exegetically accurate, instead concentrating on how philosophically interesting and helpful the definitions are.[150] Kant was used to introduce the topic of hypothetical imperative, rather than to introduce a section on Kant scholarship.

Firstly, hypothetical imperatives as I've described them are concerned with what agents *desire*, rather than with what agents *will*. Competing definitions may want to focus on willing instead of desiring. Johnson and Cureton give the following example in the SEP entry on Kant's Moral Philosophy:

[150] An example of someone who does discuss other differences between Kant's idea of a hypothetical imperative and more modern concepts is Schroeder, M. (2005).

> "if you want pastrami, try the corner deli" is […] a command in conditional form, but strictly speaking it […] fails to be a hypothetical imperative in Kant's sense since this command does not apply to us in virtue of our willing some end, but only in virtue of our desiring or wanting an end.[151]

According to Johnson and Cureton, Kant does not want us to be so broad as to include all of an agent's desires, just the ends an agent wills. They go on,

> For Kant, willing an end involves more than desiring; it requires actively choosing or committing to the end rather than merely finding oneself with a passive desire for it.[152]

If this is right, then for Kant willing is more than just to desire something, it involves further input from the agent. This isn't a distinction that I'll explore in any more detail. My thesis focuses on the relationship between desires *broadly construed* and normative concepts like reasons and oughts. It would be most beneficial for my own purposes, then, to look at a broader conception of hypothetical imperatives than the one that Kant might have had in mind. Additionally, the system of hypothetical imperatives that include a broader construal of desires would be more informative generally. If the phenomenon of imperatives can be explained in relation to more than just a subset of desires then it *should* be. If I can say something useful about a larger concept then all the better for our understanding of imperatives. This is something I look to do. So the concept of hypothetical imperatives that I'll focus on in this chapter is one that doesn't make a distinction between willing and desiring, but rather one that focuses on desires more broadly construed.[153]

Next, I'll justify my definition against a second competing feature that an alternative definition might have: that hypothetical imperatives are the imperatives that are *not* categorical imperatives. This is the idea that hypothetical imperatives can, partly, be defined as those which are not categorical imperatives, and vice-versa. This can be traced back to Kant again, when he said:

---

151 Johnson and Cureton, (2018).
152 Johnson and Cureton, (2018).
153 This is also a move that Wedgwood, (2011) and Smith, (2004) make, according to Kolodny & Brunero, (2016): "Some suggest that this focus, on intentions and beliefs about necessary means, inspired by Kant's initial discussion of hypothetical imperatives (…) is overly narrow (…). Not simply intentions, but also desires, should be considered, and not simply beliefs about necessary means, but also beliefs about non-necessary means should be considered."

> … all imperatives command *either hypothetically or categorically*. The former represent the practical necessity of a possible action as a means to attain something else which one wills (or which it is possible that one might will). The categorical imperative would be that one which represented an action as objectively necessary for itself, without any reference to another end.[154]

One of the aims of this chapter will be to explore the ways in which these two aspects of hypothetical imperatives might be in tension, and to see to what extent a meaningful concept of hypothetical imperatives (the one I've defined them as) and categorical imperatives can overlap. For this reason I am not going to try to define either hypothetical imperatives as the opposite of categorical imperatives, or vice-versa.[155]

3.1.3 Problems generating normativity

One worry about the structure of hypothetical imperatives might be that they *appear* to obligate the agents to do more than they are plausibly obligated to do. This might make it seem implausible that what we ought to do can take this form. I'll address two different forms of this worry. Firstly, I'll discuss the problem of too-many-reasons: the worry that hypothetical imperatives may mean agents have an implausibly high number of reasons to act, and implausible reasons to perform some very odd actions. Secondly I'll discuss the problem of 'bootstrapping': the worry that oughts can be implausibly brought into existence. I'll use this sub-section to explain these worries and show why they don't pose problems for the existence of a system of hypothetical imperatives.

The problem of too many reasons

Schroeder talks about the problem of 'too many reasons' in *Slaves of the Passions*. The worry is this: if an agent desires something, then there will often be many things that the agent might be able to do to bring that thing about, many of which seem to be things that the agent actually has no reason to do at all. Let's borrow a couple of examples from Schroeder to make the problem clearer. Firstly, we have Aunt Margaret:

---

[154] Kant, (2012) p.77 emphasis my own.

[155] The differences between Kant's hypothetical and categorical imperatives are also discussed by Parfit in Parfit, (2011b) Appendix H: Autonomy and Categorical Imperatives.

> Aunt Margaret wants to reconstruct the scene depicted on page 78 of the November 2001 *Martha Stewart Living* catalogue on Mars. In order to do this, she needs to construct a Mars-bound spacecraft – for no one is going to give her one. Nevertheless, intuitively, Aunt Margaret still ought not to build her Mars-bound spacecraft.[156]

It seems implausible to describe Aunt Margaret as having a reason to build a spacecraft, but if hypothetical imperatives are reason-giving, and a hypothetical imperative tells her that she should build a spacecraft in order to reconstruct a scene on Mars, then she does seem to have such a reason.

Here's a second example. Schroeder tells us, sincerely: "you have a reason to eat your car."[157] He tries to persuade us that we have this reason because it will definitely contain at least our daily dose of iron. But even if I *do* have a desire to get at least my daily dose of iron it doesn't seem like I have a reason to eat a car, even if cars are full of iron. The account of hypothetical imperatives seems, at first, to give us implausibly too many reasons; so I need to demonstrate why that's not the case.

Schroeder himself looks to answer the problem by biting the bullet (but not because it's a good source of iron). He argues that this is more plausible than it might originally seem, for several reasons. For example, the reasons are either so heavily outweighed by other reasons that they have very little weight of their own in comparison. That explains, he says, why our reason to eat a car or for Aunt Margaret to build a spaceship seems so insignificant. It's because they *are* insignificant, even though they still *are* reasons.

Another way to understand this defence is this: when we talk about reasons we are usually trying to be helpful or informative, and so that's why referring to very insignificant reasons seems unnatural. It's not that the things we're referring to aren't reasons, but that it seems odd for us to mention them. Schroeder says:

> And so we have our two-step pragmatic explanation of why we often find it unintuitive or inappropriate to say that there is a reason for someone to do something even when, in fact, there is a reason for her to do it. It yields two predictions. If I tell you that there is a reason for you to do something that there are only poor reasons for you to do, what I say will sound wrong. But – first prediction – it will sound *less* wrong if I tell you *what* the reason is, because doing so will remove the pragmatic reinforcement of

[156] Schroeder, (2004) p.84.
[157] Schroeder, (2004) p.95.

> the standing presumption that I have only relatively good reasons in mind. And second, if I then tell you that I *don't* think it is a particularly weighty reason, I should be able to cancel the presumption, and so the unintuitiveness of what I say should go down a second time.[158]

For the most part I agree with Schroeder's analysis of these kinds of case. He's right, for example, to say that we *do* have incredibly large numbers of reasons to act in a large number of ways; and, that many of these reasons weigh very little and barely feature in our deliberations (if they do at all). But I think there is a second and complementary way to answer some of these examples, including the two that I've listed here.

Firstly, let's take another look at Aunt Margaret. Although I don't want to make too many assumptions about what Aunt Margaret's hobbies, interests and skills are, I take it that we're supposed to think of Aunt Margaret as an average person, that is, someone without any special skills, resources or training. She's no more able to build a working spaceship than I am. It seems safe to assume that *nothing* Aunt Margaret is able to do will ever get her to Mars, and she certainly won't be able to get there in her own spacecraft. So Aunt Margaret's desire *to go to Mars* is simply not something that can be achieved by her taking steps towards building a spacecraft. The structure of hypothetical imperatives gives her *no reason* to do so, because those actions just aren't the kind of thing that would help her achieve those ends.[159] Let's take a look at the structure of this hypothetical imperative:

If Aunt Margaret wants to go to Mars, then she ought to build a spacecraft.

It's not the case that this imperative gives Aunt Margaret a very small and easily outweighed reason to go to Mars, one that perhaps will never, in practice, motivate her to act. Above I described a hypothetical imperative as having two parts: the conditional desire and the act. I said that hypothetical imperatives only apply to agents when the agents have that desire *and* when the act might bring about the state of affairs that's desired. That's not the case for Aunt Margaret; even though she does want to go to Mars, there's no chance that her setting about to build a spacecraft

[158] Schroeder, (2004) p.95.

[159] As I've interpreted the case of Aunt Margaret, I don't think she'd believe that she could get to Mars either. I imagine her understanding of her own skills as being fairly realistic. But to clarify, if Aunt Margaret did believe that her actions stood a chance of getting her to Mars, then she would have (a small) reason to take them.

will mean she gets there.[160] The problem is that the imperative is just wrong, the action has no bearing on the desire. There's no reason why there'd be any motivational force, or why the hypothetical imperative should apply to her.

Something similar can be said of the iron example. The key here, I think, is to take a closer look at the desire to meet the daily recommended requirement for iron. Firstly, it isn't likely to be an intrinsic desire, but rather an instrumental desire which helps the agent to achieve other things they might desire. For example, I might want to regularly eat at least my daily dose of iron so that I can give blood, so that I'm less tired during the day, or just so that I'm generally healthier and happier. We can tell that the desire to eat one's daily dose of iron is instrumental in part because (in normal cases)[161] one wouldn't desire to eat one's daily dose of iron *unless* one believes it might contribute to those other ends that are actually desired. None of these intrinsic desires will be fulfilled by the subject eating their car.

What does the instrumental desire to get one's recommended daily dose of iron look like, properly described? It's a desire for a certain state to come about, but S doesn't desire to just literally put some iron inside of them. They most likely want several other things: health, happiness, feeling less tired during the day, the ability to give blood. At an absolute minimum, what they want is to be in a state where they've ingested a healthy amount of iron. So if we take a look at the imperative:

If S wants to eat their daily dose of iron, then they ought to eat this car.

We can see that a proper understanding of what we mean by S wanting to eat their daily dose of iron shows us that none of those states could be brought about by S eating a car. Quite simply, S eating a car would kill them. It would not make them healthy, and they would not even be able to ingest the iron and still be around to tell the tale. Once again, Schroeder's problem of too-many-reasons can be explained by the fact that some of these examples of hypothetical imperatives don't apply to the agent, the acts simply aren't something that even *might* bring about the desired end.

[160] Schroeder talks about the actions 'promoting' certain desires rather than fulfilling them, which is where he runs into trouble. (Thanks to Neil Sinclair for pointing this out). But I maintain that in these specific cases, because there is *no* chance at all of the desired outcome happening as a result of these actions, then the actions do not promote the ends either.

[161] I'm sure that we can imagine more unusual cases, but the more unusual the case is the less problematic it seems to say that those cases are the ones where the agent does have a reason to eat the car.

They wouldn't be able to give any agent any reasons to act anyway, even under our large and inclusive list of reasons.

I mentioned above that eating iron was only an instrumental desire. My opponent might wonder: does this mean I'm committed to saying that instrumental desires aren't one of the many kinds of desire I'm taking into account? After all, I've been treating desires very broadly for the rest of the thesis, and indeed I argued above that we should take desires more broadly than Kant might have wanted to. So if I want to take the opposite step here, and narrow the range of desires that count, then I need to justify why that's the right step to take. This is because some instrumental desires – the desire to ingest one's daily recommended dose of iron normally being one of these – are not really desires at all. At least, they're certainly not desires for the instrumental means *beyond* being desires for the final end. The agent desires the end, but not the things which she might have to do to bring that end about. The extent to which she 'desires' the things she instrumentally desires are no more than the extent that (1) she desires the end and (2) the instrumentally desired thing might bring about that end. If either of these things go away, if she no longer desires the end or it becomes clear that the instrumentally desired state will not bring about the end, then her instrumental desire will go away too.

This isn't to say that instrumental desires can't sometimes also be intrinsic desires, things desired for their own sake. In some cases what starts out as an instrumental desire might later become a non-instrumental desire. In other cases, just thinking about something one desires instrumentally might give the agent a greater focus upon it, and cause it to become something the agent does desire for its own sake. But this doesn't seem to be the case for something so trivial as an agent's desire to ingest the recommended daily dose of iron. There's nothing particularly exciting or valuable about getting your daily iron supply, and it's not at all a likely candidate for something that might come to be desired for its own sake- its value relies entirely on other ends that might be achieved through it (such as health).

So if we take the desire for an agent to get her daily dose of iron to only be a desire when it comes alongside the other ends (such as health) then, as we saw above, there are no circumstances in Schroeder's case where eating a car will ever bring about what the agent desires, and the hypothetical imperative will never apply.

The problem of too many reasons was that the structure of hypothetical imperatives might mean that agents have an implausible number of reasons, and implausible reasons to do very odd things, all because those things might bring about ends that the agents desire. I approached this in two ways: firstly by agreeing with Schroeder's own arguments that we do, indeed, have large

numbers of reasons to do large numbers of things. We might be forgiven for thinking we don't even have some of these particularly small reasons because they're too small to be worth thinking about, or advising others about. Secondly, I argued that some of the most unintuitive cases that Schroeder describes aren't the kinds of hypothetical imperative that would actually apply to agents after all, since the acts in question would never bring about the desired state of affairs. Next I turn to the (similar) problem of bootstrapping.

The problem of bootstrapping

The problem of 'bootstrapping' is similar to the problem of too many reasons, in that it can lead us to have some seemingly implausible normative commitment.[162] Bootstrapping problems can crop up during attempts to work out what it means to be rational: whether being rational means doing what you believe you ought to do, doing what you think is a means to achieving something else you intend to do, believing things you have sufficient evidence for, believing the logical conclusions of other things you believe, etc.[163] In short, bootstrapping in this context is when we can seem to generate normativity out of something that shouldn't generate it, by following norms of rationality. I won't go into many of the arguments on what it means to be rational in this thesis but in this section I'll explore whether bootstrapping causes problems specifically for an account of hypothetical imperatives, and whether the force from hypothetical imperatives can ever generate normativity in a problematic bootstrapping way. I'll argue that the transparent nature of hypothetical imperatives, and the way they can be divided, means that they do not. In the process, I should be able to make the account of hypothetical imperatives clearer, by giving a more thorough explanation of their normative force.

I'll begin by describing an example of a bootstrapping objection. Take the following example of bootstrapping, as explained by Kiesewetter:

> …suppose you are weighing your reasons for and against two incompatible courses of actions, say getting some work done at home and watching a football match with your friends. Your reasons, we can suppose, together require you to stay home: you have to hand in important work tomorrow, and the match is not

[162] Bootstrapping is discussed, for example, in Kiesewetter, (2017) pp.81-102, Finlay, (2014) pp.50-61, Piller, (2013), Holton, (2004), Kolodny, (2005), Cheng-Guadarjo, (2014), and originally by Bratman in Bratman, (1981).

[163] This list is given by Kiesewetter in Kiesewetter, (2017) p.14-15. He says that not conforming to these could be different ways to understand irrationality, and he describes (in order) failure to follow these as Akratic irrationality, Instrumental irrationality, Doxastic akratic irrationality, and Modus ponens irrationality.

> supposed to be very promising, after all. In deliberating, you are reaching the correct conclusion that you ought to stay at home. But then you akratically decide to go watch the football match with your friends, you need to call them and ask where they are meeting. So now you intend to watch the football match, and you believe that in order to do so, you have to call your friends. […] [W]e can now detach the conclusion that you ought to call your friends. But this seems an absurd conclusion, given that we have just said that you ought *not* to meet your friends, but rather stay home. The fact that you intend to meet your friends contrary to your own recognized reasons cannot change the fact that you ought not to meet them, and thus cannot make it the case that you ought to take steps to meeting them. You cannot 'bootstrap' a decisive reason to take some means into existence, simply by intending an end you have decisive reason not to intend.[164]

Kiesewetter worries that bootstrapping like this is a problem for structural requirements of rationality. That is, he wants to argue that we ought to follow our reasons, we ought to be rational agents, but cases like the above might seem to make this view (or certain kinds of this view) implausible.

Another helpful (and shorter) description of the bootstrapping objection comes from Holton. He says,

> Forming an intention to do something surely cannot give one a reason to do it that one would not otherwise have. If it did, we could give ourselves a reason to do something just by intending to do it; and that cannot be right.[165]

This is a useful description of what was problematic about Kiesewetter's example above. Being weak-willed has made it the case that you *ought* to follow through on those actions.

Another dimension of the bootstrapping objection comes from Cheng-Guajardo, when he says "It cannot be true in general that a person ought to do whatever will bring about her end. People sometimes adopt terrible ends."[166] Kiesewetter's example above was about an agent who was weak-willed and gave in to the temptation of watching football, but there might be similar worries about agents with immoral desires. Someone might 'will into existence' the fact that they ought to cat-call a passer-by or make some inappropriate advances to someone who works for

[164] Kiesewetter, (2017) p.82.
[165] Holton, (2004) p.513.
[166] Cheng-Guajardo, (2014) p.489.

them. Describing these actions as something they 'ought to do' just because of their desires seems problematic too.

We get bootstrapping problems in cases where (1) normative language is involved and (2) that normative language is used incorrectly, to give a different kind of normative force in the conclusion than was intended in the premises. This is what's going on in the case described by Kiesewetter above. The agent here ought not to call their friends, but by simply being weak-willed enough to decide to watch the football they've made it so that they ought to call their friends.

Whether or not this is a problem for rationality, it's not a problem for hypothetical imperatives. Hypothetical imperatives don't bootstrap any new normative force into existence that wasn't there already. No more normative *oomph* goes into the ought-statement than is put in to the beginning of the imperative, because it's conditional. Take the following imperatives:

(A) If you want to watch the football match, then you ought to call your friends.
(B) If you want to get your work done, you ought not to call your friends.
(C) If you want to follow your strongest reasons, you ought not to call your friends.

The latter two imperatives here describe the action you've decided you ought to be doing, the one which will both get your work done and be the right thing to do given the balance of all of your competing reasons. All three ought statements apply to the agent to some extent, because the agent wants each of those things. The strength of the normative force behind the claim "A ought to call their friends" is only as strong as A's desire is to watch the football match. The strength of the normative force behind (B) and (C) comes from the agent's desire to do what's overall best for themselves. But just because these desires aren't the most present in the akratic mind of the agent, that doesn't mean that they don't represent the agent's strongest desires more broadly construed and hold the greater normative force than the force behind (A).

The oughts aren't contradictory, because they are not overriding or overall oughts,[167] but rather oughts that are tied to specific desires. And any intentions or desires that the agent forms won't bootstrap any new normative force into existence. The normative force comes straight from

[167] Where an overriding ought is one that overrides other oughts you have, and an overall ought is one that is the conclusion of all of your reasons.

the desires and their strength. No unintuitive weight is given to the 'ought' claims described, only the weight that comes from the attached desires.

Bootstrapping would be problematic for an account of hypothetical imperatives only if we took the ought-statements to have more force than the conditional gives them. If we took the generated imperative to have a kind of overriding, or overall, force, when that force isn't called for. But this isn't how we should understand the ought statements. When, in the situation above, it's true that the agent ought to call their friends because they want to watch the football, then it's only true that they ought to call their friends to the same extent that they want to watch the football. This is compatible with the fact that they still, overall, ought not call their friends, because their desire to watch the football (and the normative force pushing them towards calling their friends) is less than their desire to get their work done, even if it's less strongly felt at the time.

According to my argument the two claims are not contradictory because they are different kinds of ought, and neither of them are overriding oughts. Let's look again:

(A) (If you want to watch the football match), then you ought$_{(A)}$ to call your friends.
(B) (If you want to get your work done), then you ought$_{(B)}$ not to call your friends.

Ought$_{(A)}$ and ought$_{(B)}$ are simply different oughts, so it's not contradictory for them to direct the agent in different ways. The existence of different kinds of ought is something I already discussed briefly at the beginning of the chapter. If we take an ought to be a conclusion of a certain set of reasons, then there can be a multitude of different kinds of ought that are the conclusions of different sets of reasons. After all, it doesn't seem like the following involves a contradiction:

(Ar) You have a reason to call your friends.

(Br) You have a reason not to call your friends.

We have reasons to do and to not do certain things all the time. Deliberation and weighing up reasons would be a much simpler process otherwise! So why is it that (Ar) and (Br) don't contradict? Because they're different reasons. You have a reason$_{(Ar)}$ to call your friends and a

reason$_{(Br)}$ not to call your friends. The reasons are different because they relate to different sets of desires that the agent has.

Ok, so we've pushed things a step further back. We have another set of propositions that aren't contradictory:

(Ad) You desire to watch the football

(Bd) You desire not to watch the football

What is it which makes desires like these not contradict each other? Well, desires just don't contradict like that. That's an unsatisfactory and incomplete answer, but the best one that I'll have time to cover for the purposes of this thesis. For now, I'll have to make do with establishing the fact that 'contradictory' desires don't contradict (at least, in the sense that they're possible and indeed commonly both held) and arguing from there that seemingly-contradictory desire-based reasons and desire-based oughts don't contradict *because* they are based on an agent's desires and *because* those desires don't contradict either.

I've not yet said anything more about the immoral bootstrapping that I mentioned above. Here the worry was that hypothetical imperatives meant that someone *ought* to do things that follow from their immoral desires: to cat-call a stranger, for example.[168] We can now see that the ought used in this imperative isn't making any claims about whether the agent morally, or indeed *overall*, ought to take that action. It would only be a problem if the oughts are taken to be morally loaded, but they're something more modest than that.

The move I make here is similar to a move made by Ewing, where he refers to a confusion between objective and subjective oughts.[169] But Piller is sceptical of this move, as he says here:

> This move, in my view, would deny one of the presuppositions of practical thinking. The question that characterizes practical deliberation is 'What should I do?' It is not about what I should do in this sense or in that sense. Such qualified questions – should I, just considering this or that aspect, do it? – can be

[168] For another example, Finlay talks about the imperative that if Henry wants to become a famous mass-murderer then he ought to kill as many people as he can. Finlay, (2014) p.50
[169] Ewing, (1947).

> steps towards answering what seems to be the real question of practical deliberation, which uses 'should' unambiguously.[170]

Piller's criticism, then, is that responses like the one I've just given (and even the project of understanding things in terms of hypothetical imperatives generally) are misguided, because what we really want from an understanding of normativity is one that isn't qualified by conditionals and used differently in every circumstance. Indeed, if Piller was worried about Ewing's distinction between two different types of ought, then perhaps he would be even more worried about my own theory. I do more than just distinguish between subjective and objective oughts, after all. An account of hypothetical imperatives like my own allows for as many oughts as we could have desires or sets of desires!

There are two different ways to understand Piller's criticism, and I'll explain and respond to them both. Firstly, we might understand it in the following way: an account of hypothetical imperatives doesn't reflect how people use ought statements, because we're actually all trying to talk about the same concept when we say either that an agent ought to call their friends or that they ought to stay at home and work. This is the same kind of objection that Joyce uses when he talks about the way the "linguistic population" talk about 'reasons'[171] (which I discussed in Chapter 1).

Secondly, we might understand the criticism not as an appeal to common linguistic use but as an appeal to the kind of all-encompassing theory of normativity that we *should* be aiming to explain. That we, philosophers, aren't really (and shouldn't be) looking to explain normativity in terms of different phenomena but in terms of one concrete concept, one 'ought' to rule them all, as it were.

As for the first understanding of the objection, this doesn't work because the way we use 'ought' language *is* variable. It's easy to think of circumstances in which it sounds correct to advise an agent that they ought to call their friends, meaning that they ought to do it in order to satisfy their desire to watch the football. Yet it also being correct to advise the same agent *not* to call their friends, to satisfy their desire to get some work done. It sounds at least just as likely to hear someone say "well really they ought to stay home and do their work, but if they're going to watch

[170] Piller, (2013) p.613-614.
[171] Joyce, (2001) p.102.

the football anyway, then yeah they ought to call their friends" as it does to say "it doesn't matter whether they've decided to watch the football, they ought not call their friends".

Furthermore, (and this response applies to both understandings of the criticism) those two ways of understanding ought-statements *aren't* even that different. $Ought_{(A)}$ and $ought_{(b)}$ are both still different varieties of the same thing: an 'ought', a normative and hypothetical imperative. For each of the hypothetical imperatives we're appealing to some kind of desire, and using the same kind of structure. This is pretty unified already.[172] Indeed, there's a way in which it's more unified than the kinds of ought statements that Ewing appeared to be criticising, since they aren't either objective or subjective but all equally related to the desires of the agent. Our theory of oughts does seem to be unified. Unified in the sense that the 'ought' we refer to when we think about how the agent ought to call their friends is the same kind of ought that we refer to when we think about whether the agent ought not to call their friends.

Piller's objection against systems of hypothetical imperatives doesn't work, because the extent to which the oughts I'm appealing to are ambiguous is no more than the same extent to which the term is used anyway. Indeed, they are only different types of the same unified concept.

Unlike Kiesewetter or Kolodny,[173] I don't have an answer to whether we ought to be rational. But, assuming that we are rational, and that the normative force I described above does exist, then an account of hypothetical imperatives seems like a good way to understand what we ought to do.

Conclusion

The purpose of this section is to provide an account of a system of hypothetical imperatives. I began by giving the form of hypothetical imperatives and briefly explaining their history and their use. My description was simple: they are ought-statements (*imperatives*) which are conditional (*hypothetical*) on the desires of the agent(s) they apply to. They take the following form:

[172] Other responses here come from, e.g., Thomson. In *Normativity* she argues against a unified theory of normativity; "The idea that the concepts 'must', 'obligation', 'correct', and 'ought' come to pretty much the same – a smooth, warm, conceptual pudding – is just a mistake." Thomson, (2008) p.94.
[173] Of course Kiesewetter and Kolodny answer in different ways: Kiesewetter, (2017) argues that we ought to be rational, Kolodny, (2005) that we have no such reason.

If A wants X, then A ought to φ

I then spent two more sub-sections clarifying an account of imperatives of this form. Firstly I defended my definition of hypothetical imperatives against other competing features that a definition might include: defining hypothetical imperatives in relation to categorical imperatives (something I'm expressly trying to avoid) and those which involve only a subset of 'desiring' which I found to be both more restrictive and less relevant to my overall thesis.

My next job was to establish the reason-giving nature of hypothetical imperatives and tackle two problems with it: the problem of 'too many reasons' and the problem of 'bootstrapping'. The former problem was the worry characterised by Schroeder: that an account of imperatives that gives you reasons to act based on what desires you have may end up giving you an implausibly large number of reasons. I agreed with Schroeder's own response (which was to argue that we do, in reality, have an incredibly large number of reasons to act in many ways) but also added my own. I showed that some of the more worrying objections of this kind do not generate any reasons after all, because the action specified by the imperative would not contribute to bringing about the desire in question, and so the imperative does not apply to any agents.

The latter worry, the problem of 'bootstrapping', was that it would seem to generate normative claims (that is, claims about what it is we ought to do) which actually point towards things that we ought *not* to do. I dealt with this by reminding the reader that not all hypothetical imperatives are moral oughts or all-things-considered oughts, and that there can actually, plausibly, be a wide variety of 'oughts' that don't contradict because they stem from different reasons and different desires. Next, I will move on to discussing one of the rival oughts: categorical imperatives.

3.2 Categorical Imperatives

Introduction

My previous section introduced hypothetical imperatives: imperatives that are conditional on a desire (or set of desires) of a specific agent (or group of agents). My ultimate aim for this half of my thesis is to argue that all imperatives are hypothetical ones, that all 'oughts' are necessarily

related to desires. One of the main rivals for this view is an account in which some imperatives are *categorical*. Those will be the focus of this section.

The way I've defined hypothetical imperatives is fairly clear-cut, but the concept of categorical imperatives comes with a bit more baggage. That is, there are several criteria that one might use to describe exactly what it is that makes an imperative a categorical one. In each of my next five sub-sections I will discuss a different criterion. Firstly, in **3.2.1** I'll discuss the imperatives' inherent importance and dignity. In **3.2.2** I'll discuss the idea that categorical imperatives are those which apply to us in virtue of our rationality. Next in **3.2.3** I'll turn to the idea that categorical imperatives are those that apply "for their own sake". **3.2.4** will discuss their authority and inescapability, and finally, in **3.2.5** I'll turn to the most important: that categorical imperatives are those that apply to an agent regardless of their desires.

This section will do more than just take stock of different criteria that might be relevant in defining categorical imperatives. It will also analyse the relationship between them and hypothetical imperatives. As I mentioned in the last section, I've not tried to define hypothetical imperatives in relation to categorical imperatives. What I want to do instead is to see where the two concepts overlap. I'll argue that the fifth and final criterion, that of 'categoricity', is the only one that can't also apply to hypothetical imperatives. Hypothetical imperatives can still be important and dignified, they can still apply to us in virtue of our rationality, they can still apply 'for their own sake' and they can still have an inescapable authority over agents. The only thing they cannot do is to apply to an agent *regardless of their desires*. This final criterion, I will argue, is something that simply cannot be a feature of normative ought-statements, and so we are able to understand all normative oughts as being a system of hypothetical imperatives.

I'll make one final clarification before I begin the journey through five different criteria of categorical imperatives. As I discussed in the previous section, I have not ruled out the possibility of moral imperatives being hypothetical. As will become clear, many of the proponents of categorical imperatives think that moral imperatives must be categorical. Kant, for example, said about the categorical imperative that it "may be called the imperative of *morality*."[174] For this reason, a lot of my discussion in this chapter will take place in the moral field, and involve defending not just the claim that various aspects of categorical imperatives can also be found in hypothetical imperatives, but that certain aspects of *moral* imperatives are covered by them.

[174] Kant, (2012) p.80.

3.2.1 Criterion 1: Importance, dignity

The first attribute that I'll discuss is the importance and dignity that some might think of as being particular to categorical imperatives. I'll discuss what I take this to mean, and then argue that hypothetical imperatives, too, can be important and dignified. Foot, for example, says that "…in describing moral judgments as non-hypothetical - that is, categorical imperatives – [Kant] is ascribing to them a special dignity and necessity which this usage cannot give." Williams also talks about the special importance that we give to moral obligations.[175] I take the importance and dignity of moral imperatives, in particular, to be getting at a similar enough idea that I'll tackle both together.

Part of what gives categorical imperatives this special dignity is no doubt related to some of the other criteria that I'll discuss later. It may well be the case, for example, that some of my opponents think that categorical imperatives are so important and dignified *because* they apply regardless of our desires. In fact, for each of criteria 1-4 there may be a way in which that criterion is connected to the fifth: that of 'categoricity'. But, as with each of the criteria, I'll first determine whether there's anything about it *on its own* that might serve to make a categorical imperative a categorical one. And for each one, I'll demonstrate that the criterion is something that a hypothetical imperative can fulfil just as well.

Some hypothetical imperatives will be of little importance to a particular person. If I want to be a good bearded dragon owner then I should buy some crickets when I go to the pet store. I do want to be a good bearded dragon owner, the latter half of the imperative does follow (in that buying crickets would, in this particular instance, make me just that) and so I should, indeed, buy those crickets. This hypothetical imperative is important *to me* because of the value I place on being a good bearded dragon owner, the desire in question is one that means a lot to me. But for a different agent, who doesn't even own a bearded dragon, any kinds of imperative about how to be a good bearded dragon owner don't apply to her.

The bearded dragon example is a hypothetical imperative that doesn't even apply to the second agent, because she doesn't have the desire in question, and the imperative doesn't follow. It has no normative force for her (there was a more detailed discussion of this in **3.1**). But hypothetical imperatives that *do* apply to an agent will still have more or less importance. After all, we desire things to lesser or greater extents. My desire to stay in bed isn't (thankfully) as strong as

[175] Williams, 2011 p.193, and p.202-203 in particular.

my desire to be a philosopher, and my desire to fight for social justice is stronger than my desire to have a cup of coffee. The matching hypothetical imperatives for each of these also vary in strength.

Things can also be important more generally. That is, important to society, important when considering shared values, things which are deserving of the respect of many members of that society. This sounds like the kind of importance that categorical imperatives are thought to have, but it seems like hypothetical imperatives can have this wider sense of importance as well. The desires in question can be things that are desired by society generally, and things that we would think of being particularly worthy of desire: the greater good, obeying the moral law, pursuing friendships, etc.

We can already see there are several ways that we might understand the 'importance' of hypothetical imperatives. Williams describes two features of the kind of importance that we seem to give to categorical imperatives.[176] Firstly, he says,

> … if something is important in the relative sense to somebody, this does not necessarily imply that he or she thinks it is, simply, important. It may be of the greatest importance to Henry that his stamp collection be completed with a certain stamp, but even Henry may see that it is not, simply, important.[177]

Categorical imperatives, then, are more than just important in the way that I find my obligations to my bearded dragon to be important, and more so than Henry's stamp collection is important to him. Instead, they are of the kind of 'general importance' that I mentioned above.

Williams also argues that our notion of moral importance is separate from (but related to) "deliberative priority".[178] By this Williams seems to mean a kind of 'overridingness': the idea that our moral reasons or our moral obligations would necessarily and ultimately *override* competing reasons and obligations, and that what we overall had the most reason or obligation to do would be whichever of those concepts were overriding, if any were.

---

[176] Williams describes these notions of importance as applying to *moral* obligations rather than categorical ones, and he does so as part of a criticism against the way we understand morality in *Ethics and the Limits of Philosophy* (2011). But the relevant notion he criticises is the categorical nature of those moral imperatives, which he takes to be a necessary feature of morality. In **3.2.5** I'll show why we shouldn't think of it as necessary at all. For now, though, I'll take Williams to be referring not to moral imperatives generally but to specifically *categorical* moral ones.

[177] Williams, (2011) p.203.

[178] Williams, (2011) p.203.

Whatever notion of importance we use, it seems like a hypothetical imperative can accommodate it. As long as it's something that can be desired, and something that we might be able to bring about, then it's also something that can fit into the structure of a hypothetical imperative. After all, I'm not trying to argue that *all* hypothetical imperatives will have this importance, just that they can.

This all seems fairly simple so far. But to make my case maximally clear, it would perhaps be a good idea to try to understand why my opponents might have been tempted to think otherwise about hypothetical imperatives, that they can't have the same importance or have the same dignity. Foot describes a mistake that a Kantian thinker might make,

> In the *Metaphysics of Morals* [Kant] says that ethics cannot start from the ends which a man may propose to himself, since these are all "selfish."[179]

She goes on to correct this assumption,

> It will surely be allowed that quite apart from thoughts of duty a man may care about the suffering of others, having a sense of identification with them, and wanting to help if he can.[180]

What Foot seems to be pointing out here is that agents can – and do – often have desires to do what's in the best interests of other people. Subjects desire to do what's morally right because they want to do just that, not to ultimately serve their own happiness.

This mistake (and I do take it to be incorrect) is important, because it might motivate someone to see the distinction between moral and non-moral imperatives as a distinction between doing something for its own sake and doing something to fulfil a desire. If people can desire to be good, then that distinction doesn't seem to be the right way to understand the difference between moral and non-moral (or categorical and hypothetical) imperatives.

If we take a hypothetical imperative to be an imperative simply framed hypothetically then moral imperatives can be framed in this way. For example:

---

[179] Foot, (1972) p.313.
[180] Foot, (1972) p.313.

If Maz wants to improve the wellbeing of those close to them, then they should take their partner on a date.

If Rey wants to maximise happiness, then she should campaign for human rights.

If Finn wants to keep his promises, then he should bake a cake for his friend.

If one wants to be a morally good agent, then one should x.

As Foot pointed out, it's possible to desire to be a morally good agent. This is the case for every possible definition of what it is to be a morally good agent, unless it's defined in terms of a contradiction: if a morally good agent is one who *doesn't* desire to do what's good. In that case there would be no way to cash morality out in terms of a hypothetical imperative that isn't self-contradictory. But this doesn't seem like a plausible way to define it.

It's important here to remember how wide-ranging desires can be. It's possible to agree with both the idea that an agent is only ever motivated by their own desires and to agree that that same agent can be morally good. After all, what they want might want is to act in conformity with duty, or to promote the best consequences, or to be a virtuous agent, for example. Not only can moral imperatives like these be written hypothetically, but they can be salient and relevant imperatives to real agents with these kinds of desires.

It might be the case that my opponent isn't convinced that agents can have moral desires. They might be Kantians, and/or they might have a very demanding idea of morality, where the moral agent is one with impossibly perfect desires, for example. Even if this is the case, I hope it's still clear that hypothetical imperatives can have the same level of importance and dignity that categorical imperatives can. I'll also have more to say on the compatibility of an account of imperatives as hypothetical with morality in **3.3**.

3.2.2 Criterion 2: Rationality

Next on my hit-list is the second criterion: that categorical imperatives are those which apply to us in virtue of our being rational. Johnson points out the connection between Kant's categorical imperative and rationality, describing an imperative as not being categorical "… in Kant's sense, [if] it does not apply to us simply because we are rational enough to understand and act on it, or simply because we possess a rational will."[181] Recall, too, that one of the things that Kant said when describing an imperative as categorical was that it represents itself as necessary "for a will which of itself accords with reason".[182] One of the criteria that might make a categorical imperative a categorical one, then, is that it applies to agents in virtue of their being *rational* agents.

This criterion also seems to be easily compatible with an account of imperatives as hypothetical. It seems like any imperatives that apply to agents in virtue of their being rational can also be cashed out in terms of hypothetical imperatives. I'll briefly discuss some ideas of rationality here (via a discussion of what *irrationality* might look like), and the kinds of hypothetical imperatives that agents might have in virtue of being rational in this way (or not being irrational in this way).

Kiesewetter gives us some examples of what might constitute irrationality:

> (AI) Akratic irrationality: If A believes that she (herself) ought to w, and A does not intend to w, then A is irrational.
>
> (II) Instrumental irrationality: If A intends to w, and A believes that e-ing is a necessary means to w-ing, and A does not intend to e, then A is irrational.[183]

If we suppose that being rational amounts to avoiding these kinds of irrationality, then each of these are compatible with an account of hypothetical imperatives. It's possible that agents who are not irrational in these ways might have certain desires in virtue of them (and, therefore, certain hypothetical imperatives will apply to them in virtue of those desires).

---

[181] Johnson, (2014).
[182] Kant, (2012) p.78.
[183] Kiesewetter, (2017) p.14-15. He also lists two further kinds of irrationality, but these are related to beliefs rather than intention to act, so I'll leave them out.

(AI) In virtue of wanting to avoid akratic intentionality, I might have a desire to form intentions to do the things I judge that I ought to do.

Suppose I judge that I ought to get out of bed to write my philosophy. Being rational in a certain way, and having a desire like that mentioned above, the following hypothetical imperative might be said to apply to me:

> If I desire to form intentions to do the things that I judge that I ought to do, (and I judge that I ought to get out of bed and do some philosophy) then I ought to form an intention to get out of bed and do some philosophy.

The hypothetical imperative is not very pretty, but it works. We can do the same for (II).

(II) In virtue of wanting to avoid instrumental irrationality, I might have a desire to intend to do all of the necessary steps that are involved in acting out my other intentions.

Suppose, again, I believe that I ought to get out of bed to write my philosophy. I might, further, think that setting my alarm is a necessary means to getting out of bed to write my philosophy. I might have the following hypothetical imperative:

> If I desire to intend to do all of the necessary steps that are involved in acting out my other intentions, (and setting an alarm is a necessary means to achieving my other intention of getting out of bed to write some philosophy) then I ought to set my alarm.

Under both of these descriptions of rationality (or descriptions of how to avoid irrationality), it seems like we can have hypothetical imperatives in virtue of being rational in these ways.

Are there other ways to think about rationality that might not be so easy to account for with an account of oughts as hypothetical imperatives? Perhaps my opponent might think that

acting in accordance with rationality means specifically not acting (only) in accordance with our desires, but rather acting in accordance with some end regardless of whether or not we desire it. This is reminiscent of reasons externalism, as discussed in Chapter 1: the view that an agent has reasons to act regardless of what they desire. If this is the case, then my opponent would be taking 'rationality' to be inseparable from my fifth criterion: applying categorically, regardless of desire. I'll postpone discussion of this, then, to my discussion of that criterion.

None of the attributes listed so far seem to have, on their own, pinned down anything unique about categorical imperatives, beyond their being a subset of hypothetical imperatives.

3.2.3 Criterion 3: 'For their own sake'

Categorical imperatives are said to require us to perform actions 'for their own sake'. As with the other criteria, there may be multiple ways to understand what this means. In this section I'll treat this criterion as something separate from the imperatives applying regardless of desires (which I reserve for the final section), and first I'll say why this criterion might be understood as something different, and what that might be.

In Appendix F of *On What Matters (volume 2),* Parfit talks about two different ways that Kant describes the distinction between hypothetical and categorical imperatives:[184]

Distinction 1: Hypothetical imperatives require us to perform some action as a means to something else, where categorical imperatives require us to perform actions for their own sake.[185]

Distinction 2: Hypothetical imperatives require us to act in some way if that action conforms to our will,[186] where categorical imperatives require us to act regardless of what we will.

As I've been demonstrating in this section, there are several more ways in which categorical imperatives supposedly (but don't actually) stand out from hypothetical imperatives, but Parfit and

[184] Parfit, (2011b) p.652-653.
[185] Worth noting here that this is *not* a way that I've chosen to define hypothetical imperatives, see **3.1.2**.
[186] I discussed the distinction between willing and desiring in **3.1.2**.

I do agree that the two above distinctions are, themselves, distinct, and that Kant is wrong if he conflates the two. The former is the focus of this section, the last is the focus of **3.2.5**.

I should (of course) say why it is that these two understandings are distinct, why 'applying regardless of the will of the agent' and 'requiring agents to perform actions for their own sake' are different attributes that an imperative can have. After all, one might think that for an action to apply for its own sake *is* to apply regardless of what the agent desires. This would be mistaken. The latter of these simply states that categorical imperatives must apply no matter what the will of the agent is, whether they desire to be good or bad, even whether they desire anything at all. They apply without the first half of the imperative at all, without the hypothetical 'if' that refers to the agent's desire. Categorical imperatives look like this:

Shardene ought to attend the protest

Darnette ought to care for her friends

Unlike with the hypothetical imperatives there are no conditional desires here that make the imperative true. The imperative is just true regardless of desires. Shardene ought to attend the protest even if she doesn't care about the issue. Darnette ought to be kind to her friends whether or not she cares about them or her relationship with them.

But the criterion of requiring the agent to perform actions for their own sake is only about the action, not about the desires. It's still possible for an imperative to direct an agent to perform an action for its own sake, and for the imperative to apply to the agent in virtue of their desires. Take the following examples:

If Shardene wants to encourage her government to act, then she ought to attend the protest.

If Darnette wants to care for her friends, then Darnette ought to care for her friends.[187]

Although both of the actions might be valuable, the latter one (at least) seems to be intrinsically valuable: valuable for its own sake. Darnette's caring for her friends is (or so we can suppose, fill in your own intrinsic goods if you'd prefer) something good not just as a means to some other ends but good on its own.[188]

Above I gave an example of how hypothetical imperatives could direct agents to perform actions for their own sake. There might be a second way in which hypothetical imperatives can capture the same kind of idea, and that's if the desire is an intrinsic or non-derivative desire rather than an instrumental desire. Take the following examples:

If you want to be healthy, then you ought to consume your daily dose of iron.

If you want to be happy, then you ought to exercise regularly.

If you want to pursue knowledge, then you ought to listen to others.

If you want to be a morally good person, then you ought to fight for equality.

These hypothetical imperatives each direct the agent to do some action because of a desire for something intrinsically valuable (and if you think that different things to these are intrinsically valuable then you can substitute your own favourites). Although the action itself is only instrumentally valuable (as I discussed in **3.1.3**, there's not likely to be anything at all valuable about eating iron for its own sake) the purpose of the action is to satisfy an intrinsically important desire. This may be another way in which hypothetical imperatives can direct agents towards ends that are valuable for their own sake.

I've given two examples here to distinguish between two ways in which a hypothetical imperative could direct the agent to perform an action that's valuable for its own sake. That is, the antecedent could be valuable (for its own sake) on its own or both halves of the imperative could

[187] It shouldn't be a problem that the second half of the imperative matches the first half. It's still informative in that the action ought to follow because of the desire that the agent has to perform that action.
[188] For more discussion of the difference between these two, and with more reference to how Kant may have meant it, see Parfit, (2011b).

be valuable for their own sake. Hopefully it's clear that either, or both, halves of hypothetical imperatives can refer to something intrinsically valuable, something that should be done or desired for its own sake. So, there's nothing unique *yet* that picks out why categorical imperatives are anything more than a subset of hypothetical imperatives.

3.2.4 Criterion 4: Authority and Inescapability

The fourth criterion is the authoritative and inescapable nature of categorical imperatives. This is the idea that categorical imperatives have authority over us, an authority that we can't escape from or choose to ignore. Williams, for example, talks about the sense that "moral obligation is inescapable, that what I am obliged to do is what I *must* do" and "that moral obligation applies to people even if they do not want it to."[189]

Opponents of systems of hypothetical imperatives would be wrong to say that the authority and inescapability of imperatives is something that's unique to categorical imperatives. Hypothetical imperatives are, in many ways, inescapable, and come with a certain kind of authority.[190]

Hypothetical imperatives are conditional on what an agent desires, but that's not the same as being conditional on whether an agent *wants the imperative to apply to them.* Agents can no more 'escape' moral hypothetical imperatives than they can 'escape' their moral desires, the moral parts of their character. I can't stop wanting to be a good person when it's inconvenient for me. That's why we feel bad when we're tempted to do something we know is wrong, and why we'll usually at least *consider* doing what's right, even when the less good thing is so appealing.

Suppose again that Shardene has to choose between going to a protest for a good cause and staying in to have a quiet day at home. Suppose as well that the right thing to do is to go to the protest, and that she knows it, but the better thing *for her* and her happiness would be to stay at home. There are several ways we might want to describe what moral hypothetical imperatives she has here:

---

[189] Williams, (2011) p.198.

[190] Here I'm taking talk of the 'authority' of categorical imperatives to be the same thing as their inescapability. On other interpretations and in other cases you might think that the authority of categorical imperatives actually refers to something more like the concept I discussed in **3.2.1**: a certain importance or dignity. For my response to that kind of authority see, of course, **3.2.1**.

If Shardene wants to be a good person, she ought to go to the protest.

If Shardene wants to do what's right, she ought to go to the protest.[191]

If Shardene wants to help others, she ought to go to the protest.

None of these are things that Shardene can escape, because Shardene can't stop herself from wanting to be a good person, to do what's right, or to help others. This is particularly the case because of the strength of that desire in most people. It's not a whim or a short-term desire that just pops into our heads in certain situations and then goes away again, it's a long-term preference.

That's not to say it's impossible for a person to change their own desires. There are lots of things a person could do to influence their own desires: they could concentrate on certain off-putting (or tempting) thoughts, they could try to distract themselves at certain times, remind themselves regularly of the negatives or positives of certain situations, do some purposefully selective research into what things are like, etc. But long-standing desires like moral desires aren't likely to change on a whim. For many I'd say it's likely to be impossible, and for many others practically so. In the same way, for those many people moral hypothetical imperatives are inescapable.

Furthermore, agents can and do desire to meet high and authoritative standards. If we suppose that morality is a set of objective rules or principles then we can have long-standing desires to meet those standards as much as we can, even when those standards require us to act in ways that go against the desires that seem strongest to us at the time. This is another way that hypothetical imperatives can have an inescapable authority over us.

Perhaps there's another sense of 'inescapability' that I've not covered. Although I've touched on one understanding of what it would mean for an imperative to be inescapable, perhaps there's another, stronger, sense. Inescapability might not be about the ability to choose to escape, or to take action to escape. We might take inescapability to mean that it would literally and by-definition be impossible to escape, regardless of whether we could change our desires. But, as you

[191] Doing what's right, of course, isn't always the same as being a good person, unless whether you're not a good person changes every time you act. (I take that to be implausible, and rather that good people can do bad things, and that bad people can do good things. To describe someone as good or bad sounds like a judgment to make about someone's overall character or actions, not just those of a single instance.)

might have noticed, that kind of inescapability is just another way of describing my fifth and final criterion: categoricity (applying regardless of desires). This is the criterion I turn to now.

3.2.5 Criterion 5: Applying categorically

The final criterion that I'll turn to is in some ways the most significant, because it's the criterion that, by definition, cannot be shared with hypothetical imperatives. It's that the imperative applies to agents *categorically*, that is, regardless of what the agents desire.

As Foot describes, the nature of hypothetical imperatives means that the 'ought' should be withdrawn if the desire is also withdrawn. For categorical imperatives there's no such need, since these oughts are meant to apply categorically, to all agents and in all circumstances.[192] Railton, for example, says:

> To show that a norm or reason is non-hypothetical is not to show that it is utterly without condition. It is only to show that it would necessarily apply to any agent as such, regardless of her contingent personal ends.[193]

Williams also talks about this kind of categoricity, (using the language of what we 'must' do rather than my own preferred language of what we 'ought' to do). He says:

> Sometimes, of course, "must" in a practical conclusion is merely relative and means only that some course of action is needed for an end that is not at all a matter of "must." "I must go now" may well be completed "… if I am to get to the movies" where there is no suggestion that I have to go to the movies: I merely am going to the movies. We are not concerned with this, but with a "must" that is unconditional and *goes all the way down.*[194]

The kind of 'relative' must he's talking about here is a hypothetical imperative. One thing he could mean here by the normative force of 'must' being relative to a certain end is the hypothetical

[192] Foot, (1972) pp.307-308.
[193] Railton, (1997) p.58.
[194] Williams, (2011) pp.208-209.

imperative's being conditional on a certain desire, and in this case it's the desire to go to the movies. Categorical imperatives aren't contingent on or a means to fulfilling certain desires in this sense, instead they supposedly "go all the way down". When you look for the normative force behind the categorical imperative, all you find is the categorical imperative itself.[195]

From here I'll make two moves. Firstly, I'll distinguish between two ways in which something might apply categorically: weakly and strongly. I'll follow Foot in arguing that weak categoricity is something that doesn't seem to be either unique to the kinds of imperative that are normally thought to be 'categorical imperatives', such as moral imperatives, etc. Furthermore, it's not plausibly a feature of *normative* oughts, but rather something more like a description of rules. *Strong* categoricity, I will then argue, does not exist.

I'll begin by describing weak categoricity. Weak categoricity is when something like a rule (or, more relevantly, an ought!) can be said to apply to people regardless of their desires, but there's no extra force that prescribes that they follow that rule other than, say, convention. Take this example from Joyce,

> Consider Celadus the Thracian, an unwilling gladiator: he's dragged off the street, buckled into armor, and thrust into the arena. Despite his protestations, he is now a gladiator (I take it that being a gladiator is rather like being a shark attack victim – something that can be forced upon one very unwillingly). Let's imagine that there are various rules of gladiatorial combat: you ought not throw sand in your opponent's eyes, for instance. Celadus is a gladiator, subject to the rules, and so he ought not throw sand in his opponent's eyes.[196]

The rules of gladiatorial combat apply to Celadus here even though he has no desire to follow them. Take another example from Foot,

> …we find this non-hypothetical use of "should" in sentences enunciating rules of etiquette, as, for example, that an invitation in the third person should be answered in the third person, where the rule does not fail to apply to someone who has his own good reasons for ignoring this piece of nonsense, or who simply does not care about what, from the point of view of etiquette, he should do.[197]

---

[195] This idea sounds closely related to the way proponents of categorical imperatives sometimes talk about the categorical imperative directing us towards actions that are valuable for their own sake, instead of as means to another end. I discussed and dismissed this possibility in **3.2.2.**

[196] Joyce, (2001) p.34.

[197] Foot, (1972) p.308.

In both of these cases there are oughts which apply to agents regardless of what those agents might desire. Joyce refers to both of these examples as examples of a "weak categorical imperative".[198] He asks us to consider the difference between saying that the gladiator ought not to throw sand in his opponent's eyes and moral imperatives, and claims that there is some kind of extra ingredient in the moral 'strong' categorical imperatives that these weak categorical imperatives don't have.

As Foot and Joyce both demonstrate, weak categoricity isn't something that only applies to moral imperatives. The ability to apply to an agent regardless of their desires just doesn't seem to be the kind of thing which proponents of categorical imperatives really want to get at.

Regardless of its effectiveness, I've still shown that there's a kind of imperative that isn't also a kind of hypothetical imperative. Weak categoricity is a feature that hypothetical imperatives just cannot have. So what does this mean for my account of normative oughts as being conditional on desires? Thankfully, not much. Weakly categorical imperatives are not normative 'oughts'. My argument for this is similar to one of my arguments against some uses of reasons language that doesn't correspond to normative reasons in Chapter 1. For an ought to be *normative* it needs to have some kind of normative force behind it. It needs to apply to a specific agent or a group of agents, *and* hold some kind of force over them. Of course, there are some cases in which weakly categorical imperatives can also be normative, but those are the same occasions when they will also be hypothetical imperatives.

For example, suppose I say to you that you ought to eat with your mouth closed. I'm appealing to a rule of etiquette here, and etiquette is something that can apply to people regardless of what they desire. But in this case it's also a hypothetical imperative which is implicitly conditional on your desires, such as your desire to follow the rules of etiquette and your desire not to annoy me while you eat. In cases where the agents don't have the relevant desires at all, such as the case of Joyce's gladiator, then the imperative isn't a normative one at all. To say that 'Celadus ought not to throw sand in his opponent's eyes' is not really a claim about what *he* should do, or a claim about Celadus at all. Instead, it's more like a description about the rules of being a gladiator.

I said at the beginning of this chapter that 'oughts' are the conclusion of reasons. When we weigh reasons up, then we find out what we ought to do, where that ought is tethered to the same desire (or set of desires) that the reason (or set of reasons) is. A weakly categorical imperative isn't the conclusion of normative reasons, because it's not tethered to any specific agents and their

[198] Joyce, (2001) p.36.

reasons. Instead, when we talk about someone having reasons of etiquette or reasons of gladiatorial combat (and we're not talking about agents who have other reasons to follow the rules of etiquette or gladiatorial combat) then we seem to be describing something different than a normative reason. Perhaps, rather, we mean what the agent would have a reason to do *if* they desired to follow that particular set of rules. Perhaps we mean to describe what actions an advocate of those rules sets would prefer for them to take. Finlay makes a similar point:

> To be thorough we can begin by observing a range of cases that don't strictly involve categorical or prescriptive uses, though are easily mistaken for them. If the speaker is unaware that the presupposition is false then the utterance may just be a failed recommendation. She may alternatively be hoping the agent prefers the end, despite his declarations of indifference or other contrary evidence. Others' psychological attitudes can be murky to us, and their testimony unreliable. One also often encounters a rosy view of human nature that even the most reprobate monster deep down has a hidden core of "humanity", and some utterances that appear categorical may rather be optimistic attempts to appeal to this supposed inner humanity. Or a speaker may be bluffing, or deceitfully pretending there is some end the agent would relevantly prefer, which might successfully engage the agent's [other] desires[.][199]

Finlay gives us a range of examples of when we might use weakly categorical imperatives, and what use they might serve other than being a normative ought-statement. Often, it may just be the case that the speaker *hopes* that the imperative is hypothetical, and appeals to some desires of the agent in question. This is reminiscent of a similar argument made against external reasons in Chapter 1.

I have now argued that weakly categorical imperatives are either the kind of thing which can also be explained as a hypothetical imperative, or they're not normative oughts at all. But before I move on, I'll look at what it means for an imperative to be *strongly* categorical. Joyce has this to say,

> This extra ingredient is what Foot calls "the fugitive thought," and most of her paper is devoted to showing that there is no fugitive to be found. She hypothesizes that the difference between a strong, Kantian categorical imperative and a weak, institutional categorical imperative, is that the former purports to bring with it a reason for action, while the latter does not.[200]

[199] Finlay, (2014) p.185.
[200] Joyce, (2001) p.37.

Joyce thinks that strongly categorical imperatives must give the agents *normative* reasons to act. But he agrees with me that this is impossible for categorically-applying imperatives to do. Categorical imperatives that apply in this *strongly categorical* sense cannot exist, the best we can do is either weakly categorical imperatives that aren't normative, or hypothetical imperatives that are.

There is no extra ingredient that can make a categorical imperative into a strongly categorical one. I have already ruled out a number of possibilities in **sections 3.2.1-3.2.4**. All that really remains is the ability for the imperative to motivate the agent to act.[201] But, as I have discussed in the introduction, I am working with a broad enough understanding of desire as to include anything that can motivate an agent to act. Strong categoricity, then, cannot exist. There cannot be a necessarily motivating kind of categoricity that applies regardless of the agent's desires.

Conclusion

The overall aim of this chapter is to argue that all normative 'oughts' are necessarily contingent on the desires of the agent or agents in question. I began by introducing the terminology of hypothetical imperatives: ought statements that apply to agents conditionally on their desires. This section introduced a contender for a rival to a system of hypothetical imperatives: categorical imperatives. These were harder to define clearly, and I went through five different criteria that categorical imperatives have been said to have: (1) importance and dignity, (2) applying in virtue of the rationality of the agents, (3) prescribing actions to be done for their own sake, (4) authority and inescapability, and (5) applying regardless of desires.

In **sections 3.2.1-3.2.4** I demonstrated that the first four criteria can all apply to hypothetical imperatives as well as categorical ones. Hypothetical imperatives can be important and dignified because agents can and do have desires for very important ends, including moral ones. Following our hypothetical imperatives is, too, something that can be characteristic of a rational agent. Hypothetical imperatives could direct agents towards intrinsically valuable ends, and ends that were desired or valuable *for their own sake* as opposed to instrumentally. Finally they could be inescapable and authoritative, because many of our desires are involuntary. Given these four, my opponent at best showed that categorical imperatives were a *subset* of hypothetical imperatives.

[201] This kind of view is held by, for example, Aristotle, (2009), Korsgaard, (2009), Kant, (2012).

The fifth criterion required a different approach. It was that the imperatives applied to agents *regardless* of what their desires were. This, of course, meant that this criteria excluded the possibility of the imperatives being hypothetical. Instead, I argued here that the fifth criterion is just not something that can describe an actually normative 'ought', and rather these kinds of categorical imperatives are something else, such as descriptions of rules or laws.

I have one more task for this chapter. Error theorists such as Joyce, Mackie and Olson[202] agree that strongly categorical imperatives – ones that *do* have a normative force over the agents they apply to – do not exist. But they use this as evidence that morality itself is does not exist either. My final section of this chapter will demonstrate that categorical imperatives are unnecessary, and that morality *can* and *should* be seen as a system of hypothetical imperatives.

3.3 Moral Realism and Hypothetical Imperatives

Introduction

This section is about more than just some interesting upshots of an account of hypothetical imperatives; there are at least two reasons why a defence of moral realism has a particularly significant place in this chapter. Firstly, and perhaps most obviously, because it's just important to know whether morality is real. My thesis focuses on the relationship between certain normative concepts and desires, but one of the reasons that these topics are so interesting is because of the better understanding that this gives us of ethics and moral psychology: why agents do what's right, under what circumstances we can be good, and whether morality is overly demanding of us. It's important, therefore, to know whether my arguments, if correct, show moral language to be about something that doesn't exist.

Secondly, it's important to defend moral realism here in order to further defend my account of hypothetical imperatives against categorical ones. Some of my opponents might think it's more important to hold on to beliefs about moral realism than it is to hold on to desire-based normative concepts, no matter how convincing my arguments might have been otherwise. If my account shows moral realism to be false, then they might use this fact as overriding evidence against my account.

[202] Joyce, (2001), Olson, (2014) and Mackie, (1977).

This section will be made up of two sub-sections. I'll begin in **3.3.1** by explaining the threat of error theory: the non-realist moral theory that's the source of potential trouble. Then, in **3.3.2** I'll argue that the most important facts of morality aren't things we'd lose under an account based on hypothetical imperatives.

3.3.1 The threat of error theory

Moral error theory is the theory that our moral judgments and discourse are based on a mistake. It's as opposed to moral realism: theory that moral properties are real properties,[203] and when we ascribe these properties to an action or a state of affairs then we're saying something true, rather than making a mistake.

According to error theorists, we believe that there's a real morality which is the subject of our moral discourse, but such a morality doesn't exist, and so our moral discourse is false. Describing error theory, Joyce says that it "…may be characterised as the position that holds that a discourse typically is used in an assertoric manner, but those assertions by and large fail to state truths."[204] He also says,

> Certain beliefs about language are non-negotiable. When you say this isn't really the case about that, it turns out you were talking about something different the whole time and had just been making up [its] existence.[205]

For Joyce, the non-negotiable part of morality that turns out to not exist is strongly categorical imperatives: the idea not just that we can describe the rules of morality as applying to everyone but that they *do* apply to everyone, and with a *normative force* that I above described as impossible. He thinks that our concept of morality is inescapably normative in a way that is still the case *regardless* of the desires of the agents.[206]

---

[203] This is how Dancy describes it, for example, in Dancy, (1988) p.170. Wallace describes it as the idea that morality is independent of us and prior to us in some way, in Wallace, (2006) ch.4.
[204] Joyce, (2001) p.9.
[205] Joyce, (2001) p.8.
[206] I say this to contrast the sense in which I showed in **3.2.4** that hypothetical imperatives can be 'inescapable' and 'authoritative' because of the way that we cannot escape our own desires, and our desires (particularly our most strongly held ones, often including our moral desires) have a kind of authority over us. This kind of authority and inescapability is not the kind, I think, that Joyce would find problematic.

Take the following quote:

> The chemist who speaks of "phlogiston" but attributes to it all the properties we associate with oxygen, the theologian who speaks of "God" but turns out just to be talking about, say, *love*, the naturalist who speaks of "moral depravity" but leaves out any notion of its authoritative "must-not-be-doneness" – all have simply changed the subject, and are not talking about phlogiston, God, or moral depravity at all.[207]

Joyce sees this kind of strong categoricity as being as much an important part of our moral concepts (such as moral depravity)[208] as, for example, the qualities that turned out to make phlogiston distinct from oxygen; they're so important that getting rid of them leaves us with a different concept entirely, not just with a corrected version of the old concept.

Mackie's argument for error theory is partially motivated by similar worries. Brink describes one aspect of Mackie's queerness argument in these terms,

> …Mackie assumes that the realist thinks moral requirements apply to everyone, regardless of her desires or inclinations. Because he assumes that moral obligations must provide reasons for action and because he thinks that a person's reasons for action are based on her desires, Mackie rejects moral realism.[209]

Here the worry is the same as Joyce's.[210] But what arguments do they give for thinking that strong categoricity is a non-negotiable feature of morality? Not many, unfortunately. They largely take it to be an intuition that they expect their readers to share with them.

Brink does an excellent job of framing the argument, one that I'll borrow here:

> 1. To be under a moral obligation to do x, one must have reason to do x.
>
> 2. One has a reason to do x just in case x would contribute to the satisfaction of one's desires.

---

[207] Joyce, (2001) p.167-168.
[208] Joyce argues that none of our moral language makes sense without this idea of strong categoricity, because it's necessary for all of our moral discourse, not just a subset. Joyce (2001) p.175.
[209] Brink, (1989) p.51.
[210] For more discussion of this aspect of moral error theory see Mackie, (1977), Olson, (2014) and Shepski, (2008).

3. Hence, one can have a moral obligation to do x only if doing x would contribute to the satisfaction of one's desires.

4. Not everyone has the same desires.

5. Hence, there is no single set of moral requirements that applies to everyone; there will be different moral requirements that apply to different people in virtue of their different motivational sets.[211]

Joyce and Mackie would both defend the first premise. The reason-giving aspect of morality is what Joyce would call the non-negotiable element.

I, myself, have three different responses to the argument above. Firstly, I want to disagree with one understanding of (1). Agents can have weakly categorical moral obligations that apply to them just because *those are the moral rules,* and that's how they *would* apply to the agent in question. This is in the same way that we say that etiquette applies to everyone: it's more like a description of the rules of etiquette than it is a normative prescription. Such obligations would only be reason-giving (and only have normative force) in certain circumstances, where the agents have moral desires.

But secondly, if we take 'moral obligations' here to mean a normative obligation or imperative that does have some force, and that isn't just a description of the rules of morality, then I agree with (1)-(4) and have two more things to say about (5). Firstly I don't think that (5) entirely follows from (1)-(4), and if it does then I want to limit its scope. It's true that there's no single set of moral obligations that applies to everyone, because there will be people who don't desire to be good moral agents in any way. I discussed something similar to this in **3.2**, and it'll come up again later in this chapter. Babies, for example, probably don't have moral desires and neither do they have any moral obligations. The same goes for psychopaths, for whom I'm happy to concede that moral obligations only apply to in the weakly categorical sense (again, this will come up more later in the chapter). But it doesn't follow from this that different sets of moral requirements apply to a range of people. There may well be just one set of moral principles, and these apply to the (large number of) people who do desire to be good people. I argue for this in more detail in **3.3.2**.

Finally, if we take (5) to be a more meagre (but accurate) conclusion like this:

[211] Brink (1989) p.52. Brink's own response is to deny (2), but I find (2) to be plausible as I demonstrated in Chapters 1 and 2.

(5*). Hence, (normative) moral requirements do not apply to everyone, only to people with the right kind of motivational set.

…then I don't think that this claim is incompatible with moral realism at all, even though it is incompatible with what Joyce called the "non-negotiable" aspect of it: the strongly categorical nature of the imperatives, the ability to apply to everyone regardless of their desires. But I want to do more than just state my own intuitions. In the rest of this section I'll give reasons why moral realism survives being a system of hypothetical imperatives.

3.3.2 Moral realism without strong categoricity

Strong categoricity is not a necessary feature of morality. This section will be where I argue for that claim. I'll do so in two sections, and in each of these I'll discuss one or more important features of morality that might have worried my opponent, since they've been taken to exist as a part of our moral system *because* of the system's ability to generate strongly categorical imperatives. I'll show that this isn't so.

Objective principles

My first target is objective and universal moral principles. I'll begin by explaining why these are such an important part of morality, and why my opponents might worry that my system of hypothetical imperatives can't account for them. Then I'll tackle their worries in two ways. Firstly, I'll demonstrate that an account of hypothetical imperatives *can* be compatible with an objective set of moral principles, and make clear the difference between subjective principles and having normative 'oughts' only apply to agents based on their desires. Secondly, I'll remind the reader of the broad range of desires that this theory takes into account. Thus a system of hypothetical imperatives might be about "doing what you want" in a way, but certainly not about doing what you *think* or *feel* like you most want.

One aspect about morality that might seem important is that it has real principles, that it is a kind of 'higher authority' that exists independently of ourselves.[212] I won't say much in defence

[212] It's not necessary to think that there are moral principles to be a moral realist, though. Particularists, for example, (such as Dancy, (2004)) think that ethics is real but can't be formed into principles. If this is the

of this idea (indeed, if my opponent thinks that this isn't a necessary part of morality then I don't need to argue that my theory is compatible with it) but I take it to be a widely shared intuition among many moral realists. Morality, my opponents might think, needs to be in some way universal. It needs to explain why the morally best options aren't always the easiest things to do.

At first this might seem difficult for the account of hypothetical imperatives to explain. Firstly, because the idea of objective principles might seem to be in tension with the idea of grounding normativity in the desires of the agent. Secondly, because of the way that our moral duties so often conflict with what we want to do. Indeed, sometimes it seems like an action is the most morally praiseworthy and morally required precisely because it's the action that we find most difficult, that we *least* desire to do. If the morally best option is for me to give up my place on a lifeboat for someone else then it doesn't seem to be morally best in virtue of some desire I have to remain on the sinking ship. Such an act would be supererogatory, above-and-beyond the call of duty, precisely because of the sacrifices to my own future and happiness that I would have to make.[213]

A categorical imperative would give everyone reasons to adhere to that objective standard, and so it seems appealing. This is related to why Brink thought that (5) was such a problem for moral realism: being moral is about more than just following your own desires, and it doesn't seem to be about having a different set of rules for everyone. Next I'll explain how a theory of hypothetical imperatives can account for objective principles too.

The existence of strongly categorical imperatives isn't necessary for the objective moral standard itself to still exist. Morality itself can still be a set, independent standard without everyone always having reason to follow it. An analogy here might be with the law. The law of a country (we can suppose) is a defined set of rules that exist independently of that country's citizens and their desires. This doesn't mean that everyone has a reason to follow all of the laws of that country. Some laws might be unjust or irrelevant (something less likely to be the case in morality), some people might have cares, projects or other reasons that prevent the law from being reason-giving. Perhaps most significantly, not all people are even going to be citizens of that country. I don't have any special reason to follow the laws of the USA when I've never been there, for example.

---

case then I won't need to justify that an account of morality as a system of hypothetical imperatives is compatible with the existence of objective moral principles.

[213] Archer argues in Archer, (2015) that supererogatory actions are better seen as necessities than sacrifices, but I take it that this doesn't affect the point I want to make here. If need be, I can replace the language of supererogation here with just the language of the morally exceptional.

Morality is like the law in that way. It can be an objective standard, it can be independent of our desires, and it can be real, all without it needing to be the case that our reasons to follow it are also independent in that way. Suppose it turns out that what's morally good is what produces the greatest happiness. That will still be the case, even if nobody has a reason to bring about happiness. The objectivity of morality and moral principles does not depend on moral imperatives being strongly categorical.

I said above that morality can sometimes seem to be precisely *about* having an obligation to do what you really don't want to do, as with the example of the lifeboat. But it might be worth remembering here the broad range of desires that count for hypothetical imperatives, in particular that I want to include standing and long-term desires,[214] desires that the agent has but might not be the most strongly *felt* at any particular moment. Recall the example of the philosopher who can only think about their particularly vivid desire to stay in bed.[215] It would be wrong to say that that desire, despite being so painfully prominent in her thoughts at the time, is the only desire she has. She does still desire to get her work done, she desires to become a great philosopher, she desires being the kind of strong-willed person who can really power through temptation. Indeed, it might be the case that the desire to stay in bed is felt so strongly at that time precisely because of the conflict between that desire and the desires of her long-term projects and goals. She feels pulled, by desire, in both directions, and one of those directions just seems particularly painful to her at the time.

When we think about this example we can see exactly the kind of inner struggle that we think of as characteristic of the moral examples, too. Having an agent's imperatives be contingent on her desires doesn't mean that there'll never be conflict and struggle, that there'll never be tough situations, or that there'll never be times that she is compelled by moral imperatives to act against what her strongest desires *seem* to be. With morality as a system of hypothetical imperatives, the struggle is still real.

There's one more important point I want to make before I move on. An account of hypothetical imperatives doesn't imply that moral principles change with the desires of the agent, but rather the extent to which the objective principles apply to the agent[216] can change with the

---

[214] These distinctions are made by Schroeder, T. (2017). See also the discussion on desire in the introduction of this thesis.

[215] This example was borrowed from Foot, (1972).

[216] You might think it's not about an 'extent' as much as it is a yes-or-no thing, perhaps either the objective principles apply to you or they don't. Either view is compatible with the idea that what determines the success is the agent's desires; you might think that the desires give us a sliding scale or that there's a certain threshold for how strong a desire needs to be, or exactly what kind of desires are needed.

agent's desires. Now is a good time to argue why that's plausible, and I'll do so here making particular note of marginal cases.

In the legal analogy, too, there are people for whom the laws don't apply. People of a certain age, or people who live somewhere the law doesn't apply, for example. The same applies in the case of the *moral* law. There are lots of entities without moral obligations. Definitely rocks, trees and mugs of gin, for example. Animals probably don't have moral obligations either, and certainly not *most* animals. Babies don't seem to start out with any, but will usually get closer and closer to being moral agents the older they get and the more they develop (or you might think that they acquire them all at once at a certain point in their lives; either way they tend to start out with none and end up with some). An account of moral 'oughts' (and so also of moral obligations) that connects them necessarily to the agents' desires gives us an explanation as to why this is: moral oughts apply to entities if/when those entities start to have intrinsic moral desires.[217]And that answer seems to track on to all of the marginal cases in the right way. Here I'll go through some of them.

Inanimate objects don't have desires, and they don't have moral obligations. That one's fairly simple, but also not particularly informative. There are lots of other things, after all, that inanimate objects don't share with humans and which might explain why they don't have moral obligations. They also don't act, for example. Animals, then, might have desires, but they're unlikely to have the kinds of desires that would satisfy an ethical theory's conditions for being *moral* desires. Animals don't generally desire to be good in the way that we take to be morally relevant. But what if they did? If there are intelligent animals out there – apes, perhaps – who form genuine friendships or loving relationships, who are able to understand some concept of goodness that's recognisably moral and to then intrinsically desire to act in good ways or bring goodness about, then it seems like there would be a good case for those animals being subject to moral oughts. This all depends, of course, on our understanding of the psychology of animals and what

[217] Again, a better laid out explanation of this can be found in Arpaly and Schroeder, (2013). Aristotle, (2004), too, spoke about how taking pleasure and pain in the right things is something that can be brought about through moral education. This is mentioned in the Stanford Encyclopedia of Philosophy entry on Moral Character, see Homiak, (2016). Given that I've argued for the connection between pleasure / pain and desire in Chapter 2, this is compatible with my view. Schroeder agrees, saying that "Virtue … involves desiring the right things, and to the right degree." (Schroeder, 2004 p.177). For more discussion of when an agent might qualify for moral obligations see, for example, Alvarez and Littlejohn (forthcoming) who talk about a distorted capacity for moral thinking. This is compatible with my view that it's dependent on an agent having moral desires, since a lack of moral desires might do just that. This is similarly the case with Rosen (2004)'s suggestion of brain anomalies, or of being badly taught.

our preferred ethical theories take to be the right kinds of moral desires. But so far, having moral desires is a plausible criterion for moral agency.

The case might be easier to imagine with children, who regularly *do* turn into functioning moral agents. Again, according to my account, they do so not when they want to tell the truth or resist painting the walls only in order to avoid getting into trouble (because that isn't likely to qualify as a moral desire), but when they start wanting to do so just because they want to do what's right, or they understand the harms involved and want to avoid them, or because they see truth as having an intrinsic value.

Until they have these desires, when we tell them that they *ought* to do what's right, and talk to them about their moral reasons, it doesn't seem like we're using the terms in a seriously *moral* normative sense, as much as it does that we're training them. We're showing them the kinds of things that they should desire, that they should take to be important. We're describing the moral law as we see it and we keep doing it until they can see it for themselves. Until then, we'll make do with the fact that they still ought to avoid drawing on the walls because it *will* get them into trouble.

To complete the picture, we should think about what happens when humans really never do acquire the right kinds of moral desires. Firstly, it's worth saying that these sorts of people are very rare. They're who we might describe as psychopaths, sociopaths or amoralists. They're people who are psychologically set up in such a way that they cannot have the right kinds of moral desires, and cannot do what's good for the right kinds of reasons. My account here says that these people cannot be subject to moral oughts or moral reasons, and that seems right.

A further worry might be that if certain groups are just not the kinds of entities to which normative moral oughts can apply to then we cannot criticise them for moral failures. I'll address this next.

Moral condemnation

Another important feature of the morality system is moral condemnation. I'll begin by explaining why it's taken to be an important part of moral discourse, and why my opponents might worry that such criticism isn't possible on my account. I'll then respond to this worry in three parts.

Usually when we see an immoral action we condemn it, we think it worthy of our criticism. There's a moral standard (that can be universal, as just discussed) and if an agent in question fails to meet that standard then we can criticise them on that basis. It seems like something more than

just personal disapproval of that agent's actions, but a more substantive criticism that's supported by a real morality.[218] This kind of criticism might seem like an important part of moral discourse.

Moral realists might worry about the status of moral condemnation if we accept a moral system that's only made up of hypothetical imperatives, instead of strongly categorical imperatives. Strongly categorical imperatives, after all, are those that apply to agents regardless of what they desire. One of my opponent's main worries here might be that it's possible that there are agents without the relevant desires, for whom there is no moral imperative that applies to them, and requires of them that they behave properly. If moral 'oughts' only apply to people who have moral desires, then my opponent might worry that we cannot morally criticise exactly the people who are the most in need of our criticism.

Here I'll begin by conceding a point. The agent who has no desires at all to be a good person might have no reasons at all to be good. There are a range of criticisms unavailable to us about such an agent: she's not being irrational, for example, and she's not failing to follow her reasons. This seems to be the kind of criticism that Kiesewetter talks about, when he says

> Criticising someone involves more than the judgment that the criticised person has violated some standard; it also involves the judgment that the standard is authoritative for her. […] this means that the person has decisive reason to conform to this standard.[219]

He goes on,

> It seems blatantly incoherent to maintain a criticism while accepting that the person criticized had sufficient reasons for what she is criticized for.[220]

This is also the kind of objection that comes up against reasons internalism. Williams, talking about the man with no desire to do the right thing, says,

[218] Smith, Lewis and Johnston, for example, describes a "panic" at the idea that there isn't an objective rationale for morality because, for one thing, our disapproval at each other's moral views wouldn't be the same kind of serious thing we took it to be. Smith, Lewis and Johnston, (1989) p.103-104.
[219] Kiesewetter, (2017) p.25.
[220] Kiesewetter, (2017) p.29.

> There are many things I can say about or to this man: that he is ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal and many other disadvantageous things. … There is one specific thing that the external reasons theorist wants me to say, that the man has a reason to be nicer.[221]

But I still think that my account does have enough tools to deal with this problem, and I'll give three responses that defend my position.

My first response is a repeat of something I established in the previous section: that agents have a variety of desires, and what they desire most overall will often be different to what they feel most strongly at any given time. I don't stop wanting to be a productive philosopher when I'm in bed in the mornings, and I don't stop wanting to do what's right when I'm faced with an incredibly difficult moral decision. Because of this, we *can* often criticise people for failing to do what's right, and we can do so on the grounds that they're not adhering to their desires or their reasons correctly. When an agent fails to do something good we can criticise them because they're too busy paying attention to their shorter term desires over their longer term ones, choosing the easy options over those that will help them fulfil what they want the most overall.

Brink makes an important point here, following Hume. He says,

> If, for example, sympathy is, as Hume held, a deeply seated and widely shared psychological trait, then, as a matter of contingent (but "deep") psychological fact, the vast majority of people will have at least some desire to comply with what they perceive to be their moral obligations, even with those other-regarding moral obligations. Moral motivation, on such a view, can be widespread and predictable, even if it is neither necessary, nor universal, nor overriding.[222]

It seems, fortunately, like most people *do* have moral desires, even if we're often bad at acting on them, bad at prioritising them, or bad at seeing *how* to act on them. It seems like it just happens to be a fact that most people, therefore, ought to act morally, even if what they ought to do is contingent on what they desire.

Secondly, even when an agent's desires, correctly weighed, *don't* give them the most reason to do what's morally right, we can still criticise them on other grounds. Indeed, it seems like on

[221] Williams, (1995) p.39.
[222] Brink, (1989) p.49.

these occasions it would be far more appropriate to criticise them not on the grounds of irrationality but on other grounds. We can't criticise them for not following their reasons correctly, for being irrational or for failing to do what they (normatively) ought to do. (We can, of course, still say that they 'failed to do what they ought to do', but in a way that acts more like a description of moral rules or of our own preferences than a normative prescription.) But there are other significant criticisms that we can make. Perhaps most notably, we can criticise them for not having the right desires in the first place.[223]

Given that having the right desires is so crucial to (and perhaps constitutes) being a good person, it seems right that some of the strongest criticisms we make about people acting badly should be that they don't have the right kinds of desires to be morally good people. This doesn't seem to be an uncommon view; it's what we do when we call people callous, rude, or selfish. Brink makes this point,

> … [if 2 were true] Moral requirements would still apply to agents independently of their contingent and variable desires, even if they would not provide agents with reasons for action independently of their desires. Thus, we could still charge people who violate their moral obligations with immorality, even if we could not always charge them with irrationality.[224]

Foot, too, makes a similar point, when she says "The fact is that the man who rejects morality because he sees no reason to obey its rules can be convicted of villainy but not inconsistency."[225] Being convicted of villainy seems like plenty, for the limited number of situations in which that's all we can do.

This response is also relevant to other forms of criticism that might *seem* at first to be criticisms of irrationality. Take 'thoughtlessness' for an example. My opponent might worry that on my view it's difficult to criticise moral agents as being thoughtless, because it's a criticism that seems to aim at agents who have failed to notice certain things, failed to think or deliberate properly. The one kind of criticism that my system *can't* account for is to criticise people on the basis of failing to follow their own reasons when they have none of the relevant desires. But to fail to think and attend properly to certain things can be a result of having the incorrect desires. Arpaly

[223] Arpaly and Schroeder argue that to be good (and virtuous) is to have the right kinds of desires, in Arpaly and Schroeder (2013).
[224] Brink, (1989) p.75.
[225] Foot, (1972) p.310.

and Schroeder, for example, list four ways that having the wrong kinds of desire can affect cognition other than directly affecting action,

1. Through involuntary shifts in attention
2. Through changing dispositions to learn and recall
3. Through changes in subjective confidence
4. Through distortion by emotions and wishes[226]

Because our desires affect the way we learn and recall things, our confidence, our attention, etc. they affect what we do beyond just directly affecting what choices we consciously make. To criticise someone on the basis of having the wrong desires encompasses a lot.

I have one final defence of the ability to morally criticise people on a system of hypothetical imperatives. I've already argued that we can criticise agents with no (or not enough) moral desires in a number of ways, and that we can criticise people in a lot of cases for failing to follow reasons that they have. Finally, I'll say that the people for whom the last category doesn't apply – the people without the correct moral desires – are simply *not* the people who normative moral imperatives *should* apply to. This is something that I argued for more in the previous section, but is also relevant here.

When we criticise a young child for causing someone pain, we'd be *wrong* to criticise them as being irrational or for failing to follow their own reasons. It's not the case that the moral law already applied to them and they just failed to act in accordance with it. After all, they don't have the moral maturity yet to have moral reasons. And it's plausible to say, I think, that this is (at least in part) because they don't have the right kinds of moral desires yet.[227] In criticising people without adequate moral desires for being irrational we'd be making the same kind of mistake. Better to criticise them in a different way, and/or do our best to instil and encourage the right kinds of desires in them in the future.

[226] Arpaly and Schroeder, (2013) pp.223-255 in particular.
[227] This isn't to say that there aren't other relevant factors as to why the child isn't a moral agent yet. They don't just lack moral desires, but they also lack the skill or practice to recognise morally relevant facts, for example.

Conclusion

In this section I argued that two important tenets of morality still exist under a system of oughts that necessarily connects them to an agent's desires: the existence of objective, independent moral principles and the possibility of meaningful moral condemnation. I demonstrated this in two ways: firstly by showing that my account doesn't lead to moral relativism, and secondly by reminding my opponent of the breadth of desires that I want to cover in this thesis, and so moral principles will, in practice, apply to most people.

Secondly, I discussed moral condemnation. Here I argued that we get to keep the most important kinds of moral criticism – criticism about whether an agent has the right kind of character and whether they have the right kinds of desires. And in many more cases we'll be able to criticise people for failing to act rationally because those agents will be failing to pay enough consideration to their moral desires.

In a way, much of the case for the compatibility of morality with an account of hypothetical imperatives was already done in **3.2.1-3.2.4**. There I listed four features of categorical imperatives that hypothetical imperatives can share: importance and dignity, being an imperative that applies in virtue of the rationality of the agent, directing agents to do things for their own sake, and being an imperative that has authority / is inescapable. If we can get all of these things without strong categoricity, then this is even more reason to think that strongly categorical imperatives aren't a necessary feature of morality.

What I really hope to have shown in this chapter is that a system in which oughts don't apply to agents unless they have certain desires needn't look that different to how we thought morality should look anyway, and how my opponents thought moral oughts should look as categorical imperatives.

Conclusion

Chapter 3 introduced my subjective view of 'oughts', which I referred to as a system of hypothetical imperatives following Foot. According to my view, all normative oughts are contingent on the desires of the agent the ought applies to, just like with the normative reasons that came before them. I began by defending my account of hypothetical imperatives, particularly against alternative definitions, against the problem of too many reasons and against the problem of bootstrapping.

In **section 3.2** I argued that a system of hypothetical imperatives can account for oughts that are important and dignified, that apply to rational agents, that require agents to perform actions for their own sake, and that are authoritative and inescapable. This was part of a project of demonstrating that desire-based oughts can, after all, have the kinds of qualities that we expect normative oughts to have. This project continued into **section 3.3**, where I demonstrated that some important features of morality (namely objective moral principles and moral condemnation) are also compatible with a system of hypothetical imperatives.

The only feature that a hypothetical imperative cannot have is categoricity. Weak categoricity, I argued, is not the kind of thing that is normative, but more like a description of a set of rules. This argument worked similarly to the argument against external reasons I gave in **1.3.2**. Strong categoricity, with no 'extra ingredient' to give the categoricity its strength, was also ruled out.

In the next chapter I'll turn to the final rival for a system of normative oughts as hypothetical imperatives: a certain understanding of the 'overall ought'.

Chapter 4.

What We Ought To Do:
Against an 'Overall Ought'

Introduction

Chapter 4 will argue against a certain kind of 'overall ought' that would otherwise be a rival to my account of normative 'oughts' as being hypothetical.

Wider context

This chapter will be the final step in my arguments defending a subjective account of normative, deliberative concepts, and the second chapter that sets out to specifically defend a subjective view of 'oughts', that I came to call 'hypothetical imperatives'. In the previous chapter I explained what I took a hypothetical imperative to be, and argued that they can do a lot of the work that had previously been taken to be work that only a 'categorical imperative' – an ought that applies to agents regardless of what they desire – could do. I then argued that the only thing that was unique to categorical imperatives, the 'categoricity' itself, was not the kind of thing that could plausibly be part of a normative 'ought' after all.

There's one more rival 'ought' that I'll consider, and that's what I'll call an 'overall ought'. For this kind of ought the emphasis isn't on whether or not it can apply to agents regardless of what they desire, but rather on providing a balance between other existing oughts. Because these oughts themselves (the ones the overall ought aims to find a balance between) might be contingent on the agent's desires, the overall ought is a different phenomenon from the categorical imperatives. But the overall ought I'll target isn't the same as a hypothetical imperative either, because it's not itself something that the agent ought to do in order to fulfil a certain desire.

To begin my argument, **section 4.1** will look to explain in more detail what the target 'overall ought' is, and make it clear how it differs from the kind of desire-based hypothetical imperative that I've defended previously. **Section 4.2** will then house the bulk of this chapter, as I argue that such an 'overall ought' is implausible on the grounds that it would require us to agree to some unintuitive claims about supererogatory acts. Finally, in **section 4.3** I'll demonstrate more explicitly why a *different* kind of overall ought, one that's more directly related to the desires of the agent, would be the best way to escape the objections this chapter raises. This chapter is relatively brief, but (as I hope to show) contains some important contributions to the literature, firstly by giving a new argument against a commonly used overall ought concept, and secondly by demonstrating how a desire-based overall ought would escape such an objection.

4.1 What We Overall Ought to Do

Consider Helen.

> Helen accepts that, given the needs of the poor, she is morally justified in keeping for herself only what she requires to meet her basic needs. She also thinks it's important to do what is right, but she really wants to go hiking amidst spectacular mountain scenery, which involves spending money on travel, accommodation, and hiking equipment. She doesn't think that her desire to go hiking provides a moral justification for spending this money, but she also doesn't think that it is irrational for her to spend it.[228]

Helen concludes that what she morally ought to do is to give most of her money to charity, and what she ought to do based on her own self-interest is to spend it all on hiking. Helen might want to ask a further question: what ought she do overall, taking these different reasons into account?

[228] Singer, (2009) p.389.

I take the kind of overall ought the chapter targets to be widely used. Dancy, for example, discusses an 'overall ought' as what we have most reason to do given all the 'contributory reasons',[229] and Davidson and Dorsey both talk about what we should do all-things-considered.[230] Another example is Hurka, who describes the overall ought as it is used by a school of thought he calls the 'Sidgwick-Ewing' school.[231] Zimmerman discusses an overall ought from the perspective of virtue ethics.[232] But one of the best treatments of the overall ought can be found in McLeod,[233] as he argues for a coherent theory of an overall ought (which he refers to as the 'Just Plain Ought'). He frames it in terms of the kind of question above, which asks what to do when there are conflicting oughts. He then says,

> The [overall ought] is the idea of an "ought" that is not identical to any of the relative or qualified "oughts" – that is, the moral "ought," the prudential "ought," the aesthetic "ought," and so on. […] …the concept of [the overall ought] is distinct from any relative "ought" concept.[234]

I'll try to clarify this idea.

My target is a particular concept, one that people refer to when they talk about what they *overall ought* to do. There are multiple ways that people might use the term, but the concept this chapter will focus on is a concept of what we 'overall ought' to do that (i) tries to find a balance between different kinds of oughts or reasons (such as moral and prudential) and (ii) does so by appealing to an overall standard. My opponents do not always explicitly mention these desiderata, but I have laid them out to help define exactly which concept I am targeting.

Turning first to (i), Helen, introduced above, has moral reasons and prudential reasons, which we can assume come apart.[235] Similarly, the target theory assumes that moral reasons are

---

229 Dancy, (2003).
230 Davidson, (1970) and Dorsey, (2013).
231 Hurka, (2014).
232 Zimmerman, (2008).
233 McLeod, (2001).
234 McLeod, (2001) p.273.
235 You might think that prudential and moral oughts never come apart. Anscombe and later-Foot, for example, can be seen as holding that view in Anscombe, (1981) and Foot, (2001), and those who want to look even further back can find it in the likes of Aristotle in, for example, Aristotle, (2009). These accounts are not the target of the criticisms in this chapter, because they wouldn't hold the kind of overall ought view that I argue against.

not always overriding,[236] that sometimes an agent is not required to make personal sacrifices in order to do what's morally best. It appears when we think we *should* find a balance between, say, helping others and looking after our own interests.[237] According to my opponent, there's a need for something *more* to the story than the reasons themselves, something that tells us whether Helen is balancing her other reasons *correctly*, whether she's giving the *right* weight to moral versus non-moral reasons. Woollard, for example, expresses such a need in her review of Singer's original example of Helen.[238] She wants to ask a further question: what ought Helen do overall? What's the correct amount of consideration to give to these different kinds of reason?

Moral reasons and prudential reasons are two examples of what you might want to balance, but the list might also include things like hiking reasons, Catholic reasons or bearded-dragon-owner reasons.[239] That is, reasons you might have to promote or respect the values of hiking, Catholicism, or looking after your bearded dragon(s).[240] The 'overall ought' might be just a balance of the prudential and the moral, or it might, instead, be a balance of many other kinds of reason. And there will be overlap within these categories, since to a large extent they're artificial. Because Helen enjoys hiking, for example, then what we'd call her hiking reasons and prudential reasons will often overlap.

A good analogy here might be with assessing a film. When asked to choose a favourite film, or to rank a selection, someone might have some specific criterion in mind that they judge it on. I might, for example, pick the film March of the Penguins, based on the criterion of 'the film which has the most penguins'. But, perhaps more regularly, I might try to find a film which is not just the best at representing a high number of penguins but the best *overall*, given multiple criteria that I care about: lighting, good direction, theme, *and* number of penguins. (As it happens, this would still lead me to pick March of the Penguins.)

---

[236] My opponent might think that moral *obligations* or *oughts* are overriding, but not that moral reasons or the morally best actions will always result in these. Stroud uses discussion of an 'overall ought' in this way her paper on moral overridingness, Stroud, (1998)

[237] You might want to find such a balance because of concerns about moral demandingness, for example. That is, if you worry that doing the maximally best thing in a situation is too demanding, then you might appeal to a different kind of 'ought' that lets you find a balance between that demanding morally best option and options that take *your own* interests into consideration as well. See, for example, Scheffler, (1992), Benn, (2015) and McElwee (2016) on demandingness objections, Scheffler, (1994) and Norcross, (2006) for the demandingness of consequentialism, Baron, (1987) and Annas, (1984) for discussion on demandingness in Kantian ethics and Ashford, (2003) for discussion of demandingness in Scanlon's contractualism.

[238] Woollard, (2010)

[239] A similar list can be found in Broome, (2007) where he talks about normative 'requirements', which I take to be similar to the kinds of 'oughts' I discuss.

[240] The language of promoting and respecting is taken from Pettit, (1993). I use it here to show the account's neutrality in regard to different approaches to value.

Next, (ii) was the idea that the overall ought is appealing to an 'overall' standard. I don't have much to say about what this overall standard is, since I find this unclear myself. It may well be something we just have to work out through reasoning, using our judgment.[241] For my purposes in this chapter, all that matters is that the overall ought theorist is *not* just appealing to whichever desires are strongest. Such a desire-based ought, as I'll demonstrate in **4.3**, escapes the criticisms in this chapter.

For the rest of this chapter I'll use abbreviations for three actions that Helen could possibly take when deciding how to spend her time and money:

(M) The action that Helen *morally* ought to do.

(P) The action that Helen *prudentially* ought to do.

(O) The action that Helen *overall* ought to do.

We can suppose that the action denoted by (M) won't (unless circumstances are particularly unusual) involve any hiking for Helen, rather it will most likely involve her giving away a large proportion of time and effort to charitable causes. For the sake of making this less complicated, let's also suppose what would actually be the best for Helen prudentially doesn't *much* overlap with what would be morally best for her to do, rather it would involve saving up for and having a peaceful but resource-consuming hiking career.[242] Then (O) would be the action prescribed by the 'overall ought', the one which is the 'right' balance between the other two.

Now I've tried to clarify the concept I'll list four possible ways to analyse what it may mean in moral discourse:

(1) The overall ought tells an agent what is demanded of them, what they are obligated to do.

(2) The overall ought tells an agent what it is praiseworthy for them to do.

[241] This is the kind of reasoning that Crisp refers to in Crisp, (1996). It might also be what McLeod thinks the overall ought means in McLeod, (2001) when he talks about it being a standpoint of 'reason' or 'reason-as-such', and what McElwee argues we should interpret it as in McElwee, (2007) p.366.
[242] To repeat: those who don't see the moral and the overall oughts as ever coming apart are not the targets of this chapter's argument.

(3) The overall ought tells an agent what they have most reason to do.

(4) The overall ought tells an agent what the minimum that they are obligated to do is.

I've listed four possibilities here,[243] and I'll tackle them one at a time. Some of these may often overlap, such as (1) and (3), and I'll discuss that in more detail in their individual sections later. Some of these analyses may also sound more natural than others. To me, the most natural interpretation of the overall ought is (1), but after I've argued against it my opponent may want to retreat to one or more of the other options. Options (2)-(4) may seem more plausible when (1) has been ruled out. Because of this, it is important to show why every one of these analyses is problematic. In the end I will argue that the analyses which are the least vulnerable to my criticisms on the grounds of the problem of supererogation will be the *most* vulnerable to the charge that they do not sound like an 'overall ought' after all.

4.2. The Problem of Supererogation and the Overall Ought

For each of the four analyses I will argue that the overall ought faces problems. In some cases this will be because accepting that analysis of the overall ought will also mean accepting some implausible claims about supererogatory acts. In other cases, the overall ought will have had to be qualified so much (to avoid the first problem) that it no longer resembles the kind of 'overall ought' that this chapter targets.

First, I will briefly introduce the concept of supererogation in a little more detail.[244] The concept can be traced back to Urmson, who argued that moral theories needed to be able to account for a category of actions that were 'saintly' or 'heroic'.[245] Such supererogatory actions are those which are morally good (and usually exceptionally so) but are either not required or at least

[243] Bart Streumer in Streumer (2007) p.354 gives a different list of four interpretations of ought claims generally, although not specifically of the overall ought. As well as an ought meaning that the agent might have an obligation (which I've covered with (1)) and that it would be what the agent has 'most reason' to do (which I've covered with both (1) and more specifically with (3)) he also lists the interpretation that the agent might just be expected to act in that way, and the interpretation that it would be good if the agent acted in that way. I have no problem with either of those last two interpretations of the overall ought, but take them to be different from the kind that this chapter is arguing against.

[244] For some helpful discussion on supererogation generally, see, for example, Archer, (2016), Benn, (forthcoming), Horgan and Timmons, (2010) and McElwee, (2017).

[245] Urmson, (1958) p.199.

not *morally* required.[246] Those who agree that some acts are supererogatory might think that in Helen's case, action (M) is supererogatory, and that it would be good but not required of her to dedicate all of her resources to others.

For my argument to work I don't need to be committed to the existence of supererogatory acts myself. Rather, supererogatory acts are something that my opponent is committed to. This is because the concept of the overall ought I'm working with relies on the overall ought sometimes differing from the moral ought, and that would mean that the morally best option is not required of us in at least one sense.[247]

I'll now turn to the four different analyses that I listed above of what someone might mean when they use an overall ought.

4.2.1 Demand and obligation

The overall ought tells an agent what is demanded of them, what they are obligated to do.

One job for the overall ought may be to describe what it is that is demanded of an agent, what the agent is obligated to do. This may be the most immediately natural-sounding interpretation of overall ought language. After all, it tells the agent what they *ought to do* overall, so it's understandable to think that this might consist in some kind of normative obligation or demand.[248]

Under this interpretation, (O) represents what Helen is obligated to do all-things-considered; it takes into consideration not just her moral reasons to donate money to charity but it also considers how much weight she should give to these reasons versus her prudential reasons and determines the correct middle-ground between them.

---

[246] Archer refers to these conditions as 'Morally Optional' and 'Morally Better' in Archer, (2016). He talks about the acts being *morally* required, whereas other sources like Heyd, (2016) talk about the acts not being "(strictly) required". For the purposes of this chapter I take both kinds to be cases of supererogation.

[247] One might still hold that the morally best option is *morally* required, even if those moral requirements are not overriding, and they might take that to be the case even if what's morally *required* of us is not what we overall ought to do. Even given this distinction, the morally best act is still supererogatory in the sense that it is 'above and beyond' what we overall ought to do, and that's enough for the purposes of this chapter.

[248] There's a separate option according to which the overall ought is more like a minimum requirement of what an agent is obligated to do, but I'll address this separate concern under the umbrella of (4), since I think that this particular interpretation is more like (4) than it is like (1). For now, I'll take the overall ought to refer to the single act (insofar as it's possible to narrow it down that specifically) which we are obligated to do.

The problem with looking at the overall ought as what an agent is *obligated* to do is that it means that Helen is overall *obligated* not to give up *more* of her resources to charity than whatever amount (O) would involve. Giving up more than (O) requires of her is not just something she is *not obligated* to do but something that she is *obligated not to do*. Helen's moral reasons have been weighed against her desire to go hiking, and the balance of what she is required to do has been found. But it sounds very implausible to say that Helen is obligated not to do the supererogatory act. Intuitively, it would be better if Helen gave away more of her wealth. Indeed, we call it an act of supererogation precisely because it is such a great thing for someone to do.

My opponent might claim that this objection confuses what's morally better and what's overall better. Sure, they might reply, it's not plausible to describe a morally supererogatory agent as being *morally* worse, but that's not what they would need to be committed to here. All they're saying is that the behaviour of the agent in question is 'overall' worse. We might think it sounds strange, but perhaps the strangeness is just because we're generally used to thinking of these terms as being moral terms. But the overall best is already defined as being different from the morally optimal in cases like Helen's. It would sound just as strange to say that the agent taking the best option is acting *prudentially* worse, but in this particular example that would still be true.

I don't think this is the kind of move that my opponent can plausibly make. This is because although they would want to deny that agents are required or obligated to perform supererogatory actions, I don't think that they would want to say that the agent is worse, even 'overall worse', for doing so. I would be surprised if my opponent is happy to condemn such a supererogatory agent for weighing her reasons incorrectly and placing *too much* emphasis on morality.

Part of the problem with the overall ought, and why it has so much trouble with supererogatory acts, is that it seems to act like a *quasi*-moral ought. It purports to tell us what's required, but while denying that it's always better to do what's morally best. Telling Helen that overall she should give x amount of her money to charity might, at first, seems like a good way to balance her reasons, but to describe this overall ought as something that tells her the *right* balance is not plausible.

This quasi-moral status of the overall ought is also why the overall ought, specifically, is the only kind of ought that comes across the paradox of supererogation in this way. Prudential oughts, bearded-dragon-owner oughts or hiking oughts, for example, aren't vulnerable to the paradox of supererogation. Prudential oughts tell an agent to do what's best for them, bearded-dragon-owner oughts tell an agent what best to do for their bearded dragon, and hiking oughts tell an agent how best to further the ends of hiking, and these are often not going to be the same as

what's morally best. But these oughts don't disguise themselves as being anything other than what they are: oughts grounded in certain non-moral ends. To describe a morally supererogatory Helen as not fulfilling her prudential obligations doesn't seem implausible. To describe her as not fulfilling her overall obligations is uniquely problematic.

I have argued here that the overall ought is implausible under description (1), because it requires agents to not act supererogatorily. As I said earlier, (1) is possibly the most natural-sounding way to understand the overall ought. But now that I've raised an objection with it, my opponent may want to adopt a different understanding of the overall ought to try to avoid the problem. Some of these may seem less intuitively like 'oughts', but they may still be the most charitable interpretation of what my opponents might mean with overall ought language, given the problems with (1). Next I'll look at three routes they might take. Firstly, they might want to retain an overall ought but without as much normative *oomph,* that is, without the same kind of demand / obligation on the agent. They could do this by understanding the overall ought as only being a description of what we have most reason to do, which I'll address in section (3) only after I've first set aside (2): the possibility that we analyse the overall ought in terms of praiseworthiness. I'll tackle the latter first because my response to it is so similar to my objection to (1), and so it follows more naturally. Then, in (4), I'll address a third potential escape route: understanding the overall ought not as an obligation to perform a single action, but as an obligation to perform one of a range of actions. This is analysis (4): as setting a minimum requirement for what agents are obligated to do.

4.2.2 Praiseworthiness

The overall ought tells an agent what it is most praiseworthy for them to do.

We might take the overall ought to direct agents towards what it's *praiseworthy* for them to do. Suppose Helen is having trouble deciding what to do, and she turns to ask what she ought to do. "Overall, you ought to (O)" we might reply. It seems plausible to think that Helen isn't under an *obligation* to act in that way, but rather we're just telling her what option would be the most praiseworthy.

This comes across a similar problem to that raised in (1), and correspondingly my response will be shorter. I criticised the first analysis of the overall ought on the grounds that it entailed implausible claims about supererogatory acts: that agents were overall obligated *not* to perform them. The same criticism can be made of interpretation (2); it would describe (O) as being the

most praiseworthy act, and (M) (or even some middle-ground between the two) would therefore be described as *less praiseworthy*. This is still implausible, and would make for a difficult bullet to bite.

Perhaps my opponent might find (2) appealing because of arguments like that of Wolf in 'Moral Saints'. She argues that the kind of person who does perform supererogatory actions all the time is not really a very appealing kind of person either to be or to befriend.[249] This might give my opponent reason to think someone more balanced, likely to perform the overall actions like (O) rather than to go all out towards (M), might be more deserving of praise. But this isn't so. Take this quote from Wolf,

> Despite my claim that all-consuming moral saintliness is not a particularly healthy and desirable ideal, it seems perverse to insist that, were moral saints to exist, they would not, in their way, be remarkably noble and admirable figures. Despite my conviction that it is as rational and as good for a person to take Katherine Hepburn or Jane Austen as her role model as Mother Theresa, it would be absurd to deny that Mother Theresa is a morally better person.[250]

The opponent who wants to use the overall ought to signal praiseworthiness would have to bite a bullet in which an idealised Mother Theresa is overall *less* praiseworthy for giving extra weight to her moral reasons. And although Wolf uses the phrase 'morally better' rather than 'overall better', we can see (and saw in (1)) that the point still stands.

Even if my opponent thinks that Wolf's characterisation of moral saints is persuasive, this is *still* not enough to save this interpretation of the overall ought. My objection doesn't rely on agents being complete moral saints like the ones Wolf describes. All I need for my objection to work is for an agent to be able to perform an even *slightly* morally better action than (O). Such an agent wouldn't be subject to Wolf's cutting verdict on the character of moral saints, but my opponent would still have to accept that they are (slightly) less praiseworthy for not doing (O). This still seems implausible.

I've argued against interpretation (2) in a similar way to my argument against interpretation (1). Here my opponent has the same two options: accept implausible claims about agents who act supererogatorily, or not accept this interpretation of what the overall ought means.

---

[249] Wolf, (1982).
[250] Wolf, (1982) p.432.

4.2.3 Most reason

The overall ought tells an agent what they have most reason to do.

Another suggestion for what the overall ought may indicate is that it might simply be what the agent has most reason to do. Although I have already defined the overall ought as a balance of an agent's reasons, analysis (3) understands it as only a *description* of what the agent has most reason to do, and is silent on whether we are also obligated to follow those reasons, as we saw with interpretation (1).

My criticism here begins with considering whether we can understand the overall ought as (3) without the obligation that came with (1), or if the two necessarily go together. Let's suppose first that (1) doesn't necessarily apply. Here my opponent comes across a new problem: (O) might be what an agent has most reason to do, but unless telling them so also comes with the kind of obligatory force discussed in (1) then it doesn't sound much like an overall ought. The kind of overall ought that I think my opponent is after, and the kind that I established in **section 4.1**, needs to be something more than just a simple description of weighed reasons. That is, it needs to carry some kind of normative weight that to some extent *obligates* the agent to actually follow those reasons. Overall ought language is prescriptive, but without the normative force it's only *de*scriptive. An understanding of the overall ought without being prescriptive tells the agent what the balance of their reasons is, but not that they should then do the thing that they have most reason to do. And if my opponent chooses to interpret the overall ought as both (3) *and* (1), then it is subject to exactly the same problems already discussed.

There are two moves my opponent might want to make here. Firstly, they might argue that the normative force just comes from the fact that (O) is whatever the agent has most reason to do. People are just obligated to do what they have most reason to do, that's part of what it is to be a rational agent.[251] The agent has a rational reason to balance her prudential, moral and other reasons and so she does, and rationally she's then required to perform that action. But if my argument against (1) was convincing, then it still applies if we describe the overall ought as an ought that comes from rationality.[252] My opponent would have to accept that morally

[251] For more detailed discussion on these kinds of questions of rationality, see for example Kiesewetter (2017)

[252] Wallace argues for a similar point: that "[w]e can perhaps say what it is rational to do with an eye to morality, and what it is rational to do with an eye to an individual's good, but there seems to be no common

supererogatory agents are doing something rationally wrong, they're failing to fulfil their obligations to rationality, failing to do what they have most reason to do.

Suppose Helen gives away more of her resources to charity than (O) required of her. There would seem to be a tension in how we evaluate her action. We think that she has morally performed well, but rationally performed poorly. She has done what's right in terms of morality, and failed to do what's right according to rationality. Even though 'rightness' here tracks two different things, the way we evaluate rational actions and moral actions often coincides. If rationality and morality come apart, I'm not convinced that talk of what we 'overall ought to do' tracks the former. Of course, if my opponent does find this convincing, I'm happy at least to have clarified what the quasi-moral 'overall ought' really refers to.

The other move my opponent might want to make is to appeal to a normative force *less* than an obligation which directs the agent to follow their reasons. Because it is weaker than an obligation or a demand, my opponent might say, it escapes much of the force that comes with the supererogation objection. Perhaps the overall ought *recommends* (O), rather than demands it, but in such a way that agents aren't condemnable for failing to do it. But this response fails, because the problem of supererogation comes with any level of normative force. It still seems implausible to overall criticise a moral saint on any level. Furthermore, this leads my opponent to a dilemma. The stronger the normative force to do what's overall best, the more obvious the problem of supererogation is. But the weaker the normative force is, the less like an 'overall ought' the overall ought sounds like. We use overall oughts to advise, to prescribe action, and that sounds most plausible with more normative force, more 'shouldness', behind it.

In this section I've argued that we don't use the 'overall ought' to mean what we have most reason to do without also including a kind of normative force that I argued against in (1). This time my opponent has two options. Firstly they could accept (3) in conjunction with (1) and, along with it, the implausible claims about supererogation. Their second option is to understand (3) on its own, as a simply descriptive claim that tells the agent facts about their reasons but not whether they should act on those reasons. This also seems implausible as a description of the overall ought.

4.2.4 Minimum requirement

The overall ought tells an agent what the minimum that they are obligated to do is.

currency in terms of which to cash out claims about what it is most rational to do overall." Wallace, (2006) p.131.

Finally I'll discuss the interpretation of the overall ought on which it doesn't pick out the only act that you're obligated to do, but picks out what the minimum act is that the agent ought to do. Here, we can understand it as Helen being advised that "overall, you ought to (O). You could donate more of your resources to charity than that, but (O) is the least you should do."

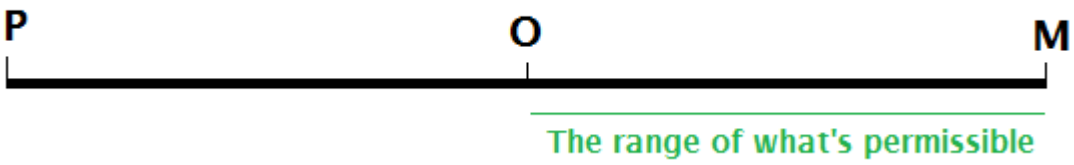

Figure 1

On this interpretation of the overall ought, the agent is obligated to do one of a range of permissible actions. They are permitted to do anything between (O) and (M). Let's examine how the range of permissible acts between (O) and (M) can, themselves, be interpreted. We have a range of acts that it is permissible (or 'overall allowed') for the agent to do, (see Figure 1) which range from (O) to (M). There are three possibilities:

(a) The agent should do something more like (O), but everything in the range is permissible.

(b) The agent should do something more like (M), but everything in the range is permissible.

(c) All options on the scale have equal merit. The agent should do any option between (O) and (M), it doesn't (overall) matter which.

None of these three look promising. (a) is the easiest to dismiss since it stumbles across the same problem that we've already encountered in (1) and then seen repeated throughout multiple interpretations of the overall ought. As I've argued above, it's implausible for an 'overall' normative theory to prefer the agents who do not act supererogatorily over those who do. Option (a), in having a preference for acts like (O) rather than (M), comes across these same problems.

What about (b), in which it is overall preferable for the agent to perform acts that are more like (M)? This time we have a sliding scale that makes sense in terms of morally exemplary agents, since the more that the agent pays attention to her moral reasons the better the agent has done. The problem here is that it doesn't look much like the *overall* ought that my opponent started out with. Instead, it looks like a regular moral ought, with a line drawn at (O). Firstly, my opponent would need to accept that when they describe what someone like Helen overall ought to do, what they *actually* mean is that Helen ought to *at least* perform (O) but preferably pay even more attention to her moral reasons. That is, when they say Helen overall ought to (O), they really mean it would be overall best if she did (M) instead. This doesn't sound like a plausible way to hear the description of what Helen overall ought to do.

Finally I'll turn to (c): the possibility that all options on the scale between (O) and (M) have overall equal merit. Again, this interpretation doesn't seem like the kind of thing people mean when they talk about what they overall ought to do, and this time it's because it doesn't actually direct the agent to act in a certain way. We might weigh up Helen's different options and advise her that she ought to allow herself the resources to go hiking once every couple of months. This doesn't sound like we're actually pointing her to a wide range of actions; it sounds like we've decided what the best option is, taking all of her reasons into account.

Interpretation (4) of the overall ought was the most complex, and gave my opponent several possible answers. It was the idea that the overall ought told the agent what was the minimum amount required of them, and any action between this and the morally best option would be permissible. With this in mind, I separated three ways to understand that claim into (a), (b) and (c). If my opponent wanted to accept (a), they would have to accept the implausible claims about supererogation that have been the basis of my main argument against the overall ought. For (b), my opponent would also need to accept that the overall ought is just a moral ought, and that when they tell an agent what she must overall do, they actually mean it would be overall better for them to do something *other than* the act they've just prescribed; better for the agent to do (M) than (O). Finally, for (c), my opponent would have to accept that when we tell an agent what they overall ought to do then we're not actually telling them to do that particular act, but rather a larger range of options. Not very informative after all. It seems, then, that no analysis of this kind of 'overall ought' language is unproblematic.

4.3 Desire-Based Overall Oughts

In **4.2** I argued that a certain kind of overall ought concept is implausible, and that my opponent would have to bite some unappetising bullets in order to carry on using the concept as it was. But this doesn't mean that we should have to give up on all overall ought language. There is a better alternative way to talk about what agents overall ought to do, and that's a way that does so by appealing to the agent's desires. In this section I'll briefly explain what such an overall ought would look like, and then demonstrate why it escapes the objections of this chapter.

A desire-based overall ought

Helen has morally good desires: desires to be a good member of society, desires to help others, whatever those desires might be. Helen also desires to pursue hiking. One way to consider what Helen ought to do overall is to do so by appealing to those desires.

Suppose we're sat down with Helen to help her to figure out what she ought to do. We ask her questions like: what does she value more, doing what's good or pursuing her hobby? Which of these things conform best with her other desires, such as to be happy, or to be fulfilled? What about her meta-desires, which of her desires does she value more?[253] We might consider what she wants to achieve compared to some external standards, like that of society or her friends and family.

Considering matters in this way might lead us to figure out what she overall ought to do in order to best fulfil the desires that she has. We might decide that overall she ought to donate, say, 30% of her time and money to charitable causes, because, we discover, she wants to do more for good causes than the people around her but without doing so much that she finds herself exhausted. This would be a good way to satisfy her strongest desires in a way that frustrates her other desires as little as possible. This kind of overall ought, then, doesn't just try to find a balance between what she ought to do prudentially and what she ought to do morally, but it find a balance that, itself, is based on her desires. Not because of some quasi-moral and unspecified standard of 'rightness', but because of what she wants to do and the kind of person she wants to be.

[253] Unless everyone present has spent too much time in academic philosophy we'd be unlikely to phrase the questions like this, but I don't think this is too far away from the kinds of things we'd aim at finding out.

Of course, agents like Helen will almost always (if not always) have conflicting desires. They'll desire to do multiple things that cannot all be done, that are exclusive of each other. It'll be difficult to figure out how to balance the desires against each other, and it might even be the case that some desires are incommensurable and cannot be weighed against one another. In such situations it will be difficult to determine what we overall ought to do given our desires. But this isn't a criticism of my account. It's a fact about how difficult it really is to determine what we overall ought to do, if there is any such thing at all.

How it escapes criticisms

The desire-based overall ought escapes the problem of supererogation, and in this section I'll demonstrate how.

The problem of supererogation for the overall ought was that most understandings of the overall ought led to implausible claims about agents who acted supererogatorily, by going above and beyond what they overall ought to have done and doing something closer to (M), what was morally best for them to do. Suppose Helen does this, after we concluded that she ought to give (O) amount of her time and resources to charity because of what she desired.

In some cases, it's worth pointing out, her actions might still have reflected what she overall ought to have done according to her desires. After all, perhaps we underestimated how much she wanted to do what's good, and how much she valued those own desires of hers. But that won't always be the case. We're not always going to do what we overall ought to do even with a desire-based understanding of the overall ought. Just as we often act imprudently by paying more attention to our weaker but more vivid desires instead of our stronger, long-term desires (like the philosopher who doesn't want to get out of bed) so too it's possible that we might give too much weight to our moral desires, compared to what the overall balance of our desires turns out to be.

The reason that this isn't a problem for the desire-based overall ought is that it's not masquerading as a standard of rightness. Just as I argued in **4.2.1**, prudential oughts and bearded-dragon-owner oughts escape the problem of supererogation because they never pretend to appeal to some quasi-moral standard. Prudential oughts are simply what you ought to do to be prudent, bearded-dragon-owner oughts to be a good bearded dragon owner. The desire-based overall ought is just what you ought to do given your desires overall, however you choose to weigh them.

Sometimes we won't have strong enough desires to do what's good, and then it just won't be the case that we overall ought to do what's good. But we still *morally* ought to do what's good.

The most plausible way to understand the overall ought, I think, is as a desire-based ought. This means that it can't purport to tell us what the *right* balance is between our prudential and moral obligations beyond determining which best fits our desires. But that's just not something that an overall ought can achieve, and that's a more plausible option than having to bite the bullets from **4.2**.

Conclusion

The target of this chapter was a quasi-moral overall ought that aims to tell us how to balance our prudential and moral reasons. I argued that commitment to such an overall ought is not plausible. The overall ought theorist, I've argued, runs into trouble when it comes to describing supererogatory agents. One route my opponent might have taken was to understand the overall ought as carrying less normative force, and so avoiding the objection this way. But the less force the overall ought carries, the more other problems stand out, such as not being able to do the job they it's supposed to, to tell Helen how much consideration to pay to her moral reasons compared to her prudential reasons.

The best way to hold onto use of overall ought language is, I argued, to understand the overall ought as simply a way of weighing up our desires. This makes sense given the rest of the arguments in my thesis, as I've argued that normative 'oughts' and reasons generally are contingent on the desires of the agent.

Conclusion

This thesis argued that what we have reason to do and what we ought to do are both contingent on what we desire. Without having desires that either *will* or *might* be satisfied by an action, then we have no reason to do that action and it's not the case that we ought to perform it.

I argued for this subjective account of normativity directly, and I also argued for it by defending it against objections. I'll use my conclusion to summarise each of these in turn.

Positive Arguments

The first argument I gave was in Chapter 1, in favour of reasons internalism. This was an argument which has appeared in the literature from several sources (Williams, Manne, Goldman and Markovits, to name a few) and with several approaches. Broadly speaking, the argument is that an agent's normative reasons to act must be reasons *for that agent*, and they must therefore in some way be the kind of reason that that agent could act for, whether or not the reasons actually motivate in practice. The best explanation of this is that reasons must in some way relate to the psychology of the agent, and in particular to her desires broadly construed.

Chapter 1 also argued against a rival view, on which agents can have normative external reasons. This is because when people make reasons-statements about an agent, and that agent doesn't have the relevant desires, one of two things seems to be happening: the reasons-statement is a mistake (the speaker perhaps hopes the agent has a different set of desires than they actually have) or they're doing something that seems qualitatively different to talking about a normative reason for that agent to act, such as expressing their own disapproval. This theme also came back in **3.3.2** when I discussed moral condemnation.

I developed my case in favour of my subjective accounts of normativity in Chapter 3, where I argued in favour of normative oughts as being conditional on the desires of the agent. I referred to this view as a system of hypothetical imperatives, after a paper by Foot. I argued over the course of the chapter that for any quality that 'oughts' are taken to have, either hypothetical imperatives have this quality or the quality itself is not something a normative ought should have after all.

Furthermore, I demonstrated in this chapter that understanding what we ought to do as a system of hypothetical imperatives is plausible for two additional reasons: firstly in **3.1.3** I showed that it can provide a clear account of where these normative concepts get their normative force from (the desires provide the force), and secondly, in **3.2.5**, that it can give us a plausible account of when a subject can qualify as a moral agent with moral obligations and moral reasons (when they have moral desires).

My final positive argument came in Chapter 4. Here I argued that when we consider what an agent ought to do *overall*, the best way to understand such an overall ought is as one that tries to balance the agent's various considerations by appealing to her desires. Other understandings, I argued, were implausible on their own terms.

Responses to Criticisms

I also used my thesis to counter the best potential objections, and now I'll briefly run through how my responses went.

In Chapter 1 I introduced the worry that reasons internalism cannot justify why an agent's normative reasons should rely on an idealised version of their beliefs but not an idealised version of their desires. There are two responses available to the reasons internalist here: either they can accept that neither beliefs nor desires should be idealised, or they can argue that the best way to determine what should be taken into account is through thinking about how easy it would be to persuade the agent of the idealised reason. This would still be compatible with reasons internalism because it would never be as easy to change an agent's desires (broadly construed) as it would their beliefs.

Chapter 2 defended reasons internalism (and value subjectivism) against two objections from Parfit. The first objection was that an agent's reasons might sometimes be contingent on their 'hedonic likings' instead of their desires. I argued that 'hedonic likings', that is, what it is for an experience to be a pleasurable one for the agent, are for the agent to have a certain kind of desire. It's therefore not possible for an agent's hedonic likings to come apart from their desire, and the objection fails.

Secondly I responded, following Street, to Parfit's 'agony argument'. This was the objection that it might be possible for subjects to have strange sets of desires such that they, for example, don't desire to avoid future agony. I argued that subjects will nearly always desire to avoid future

agony, but that if they don't, they would be such strange subjects that it's not an implausible bullet for the reasons internalist to bite.

Finally, the other main objections that I dismissed were in the beginning of Chapter 3. Here I first tackled the problem of too many reasons. This was the problem that an account on which what we ought to do follows from our desires would lead us to have an implausibly high number of reasons. I followed Schroeder in arguing that this is actually an accurate picture of how our reasons work, but I also went further than him and showed that many of the bullets he bit were unnecessary, as they were not cases in which the actions had the right kinds of relations to the agent's desires.

The 'bootstrapping objection' was the final main objection. This was the problem that a system of hypothetical imperatives would generate normativity in implausible ways. For example, agents wanting to act in ways that go against some of their stronger desires would (according to the objection) mean that they *ought* to act against their stronger desires. I argued that this objection rested on confusing certain kinds of oughts with overriding oughts. When the oughts are cleared up, the objection goes away.

I have argued in favour of a subjective view of normative reasons and normative oughts. The view not only follows from a certain understanding of desires and normativity, but is consistent with a lot of plausible views about what we have reason to do, what we ought to do, and morality.

Bibliography

Alston, W. (1967) 'Pleasure' in P. Edwards (ed.), *The Encyclopedia of Philosophy* pp. 341–347. New York: Macmillan Publishing Co. & The Free Press

Alvarez, M. (2009) 'How Many Kinds of Reasons?' in *Philosophical Explorations* 12 (2) pp.181-193

Alvarez, M., Littlejohn, C. (forthcoming) 'When Ignorance is No Excuse' in Philip Robichaud & Jan Willem (eds.), *Responsibility – The Epistemic Condition* Oxford: Oxford University Press.

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders: DSM 5 (5th edition).* Arlington: American Psychiatric Association

Annas, J. (1984) 'Personal Love and Kantian Ethics in Effi Briest' in *Philosophy and Literature* pp.15-31

Anscombe, G. (1981) 'Modern Moral Philosophy' in *Ethics, Religion and Politics: Collected Philosophical Papers Volume III,* Oxford: Basil Blackwell

Archer, A. (2015) 'Saints, Heroes and Moral Necessity' in *Royal Institute of Philosophy Supplementary Volume*

Archer, A. (2016) 'Supererogation, Sacrifice and the Limits of Duty' in *Southern Journal of Philosophy* 54 (3)

Aristotle (2004) *The Metaphysics*, London: Penguin Books

Aristotle (2009) *The Nichomachean Ethics: Oxford World's Classics*, Oxford: Oxford University Press

Arpaly, N., Schroeder, T. (2013) *In Praise of Desire*, Oxford: Oxford University Press

Ashford, E. (2003) 'The Demandingness of Scanlon's Contractualism' in *Ethics*, 113 (2) pp.273-302

Aydede, M. (2013) 'Pain' in *Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), Available online at:
<https://plato.stanford.edu/archives/spr2013/entries/pain/>

Aydede, M. (forthcoming a) 'A Contemporary Account of Sensory Pleasure' in *Pleasure: A History* Lisa Shapiro (ed.) Oxford: Oxford University Press

Aydede, M. (forthcoming b) 'Defending the IASP Definition of Pain' in *The Monist*

Baron, M. (1987) 'Kantian Ethics and Supererogation' in *Journal of Philosophy*, .84 (5) pp.237-262

Benn, C. (2015) 'Over-Demandingness Objections and Supererogation' in *The Limits of Moral Obligation,* Marcel van Ackeren and Michael Kühler (eds.) New York: Routledge

Benn, C. (forthcoming) 'Supererogation, Optionality and Cost' in *Philosophical Studies*

Bennett, J. (1997) *The Act Itself*, Oxford: Oxford University Press

Bramble, B. (2013) 'The Distinctive Feeling Theory of Pleasure' in *Philosophical Studies* 162 (2) pp.201-217

Brandt, R. (1992) 'Two Concepts of Utility' in *Morality, Utilitarianism, and Rights*, Cambridge: Cambridge University Press, pp. 196–212

Brandt, R. (1979) *A Theory of The Good and The Right*, Oxford: Oxford University Press

Bratman, M. (1981) 'Intention and Means-End Reasoning' in *Philosophical Review* 90 (2) pp.252-265

Brink, D. (1989) *Moral Realism and the Foundations of Ethics*, New York: Cambridge University Press

Broome, J. & Parfit, D. (1997) 'Reasons and Motivation' in *Proceedings of the Aristotelian Society, Supplementary Volumes* 71 pp.99-146

Broome, J. (2007) Requirements, Available online at: <http://www.fil.lu.se/hommageawlodek/site/papper/BroomeJohn.pdf>

Brunero, J. (forthcoming) 'Recent Work on Internal and External Reasons' in *American Philosophical Quarterly*

Cheng-Guajardo, L. (2014) 'The Normative Requirement of Means-End Rationality and Modest Bootstrapping' in *Ethical Theory and Moral Practice* 17 (3) pp.487-503

Corns, J. (2014) 'Unpleasantness, Motivational *Oomph*, and Painfulness' in *Mind and Language*, 29 (2) pp.238-254

Crisp, R. (1996) 'The Duality of Practical Reason' *in Proceedings of the Aristotelian Society* 96 (1) pp.53-73

Dancy, J. (1988) 'Supererogation and Moral Realism' in *Human Agency*, J. Dancy and others (ed.) Stanford: Stanford University Press

Dancy, J. (2003) 'What Do Reasons Do?' in *Southern Journal of Philosophy vol.XLI*

Dancy, J. (2004a) *Ethics Without Principles*, Oxford: Oxford University Press

Davidson, D. (1963) 'Actions, Reasons and Causes' in *Journal of Philosophy* 60 (23) pp.683-700

Davidson, D. (1970) 'How is Weakness of the Will Possible?' in *Moral Concepts*, Joel Feinburg (ed.) Oxford: Oxford University Press

Dorsey, D. (2013) 'The Supererogatory and How to Accommodate It' in *Utilitas* 25 (3) pp.355-382

Ewing, A. (1947) *The Definition of Good,* New York: Macmillan

Feldman, F. (2006) *Pleasure and the Good Life*, Oxford: Oxford University Press

Finlay, S. (2009) 'The Obscurity of Internal Reasons' in *Philosophers' Imprint*, (9) 7

Finlay, S. (2014) *A Confusion of Tongues*, Oxford: Oxford University Press

Foot, P. (1972) 'Morality As A System Of Hypothetical Imperatives', in *The Philosophical Review* 81 pp.305-316

Foot, P. (2001) *Natural Goodness*, Oxford: Oxford University Press

Goldman, A. (2009) *Reasons From Within,* Oxford: Oxford University Press

Gowans, Chris, "Moral Relativism", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), Available online at: <https://plato.stanford.edu/archives/win2016/entries/moral-relativism/>.

Gregory, A. (2012) 'Changing Direction on Direction of Fit' in *Ethical Theory and Moral Practice*, 15 (5) pp.603-614

Gregory, A. (2014) 'A Very Good Reason to Reject the Buck-Passing Account' in *Australasian Journal of Philosophy,* 92 (2) pp.287-303

Haybron, D. (2008) *The Pursuit of Unhappiness,* New York: Oxford University Press

Heathwood, C. (2011) 'Desire-Based Theories of Reasons, Pleasure and Welfare' in *Oxford Studies in Metaethics,* Russ Shafer-Landau (ed.) 6 pp.79-101

Heathwood, C. (2007) 'The Reduction of Sensory Pleasure to Desire' in *Philosophical Studies* 113 (1*)* pp.23-44

Heuer, U. (2010) 'Beyond Wrong Reasons: The Buck-Passing Account of Value' in Michael Brady (ed.) *New Waves in Metaethics*, Basingstoke: Palgrave Mamillan

Heyd, D. (2016) 'Supererogation' in *Stanford Encyclopedia of Philosophy (Spring 2016 Edition),* Edward N. Zalta (ed.) Available online at:
<http://plato.stanford.edu/archives/spr2016/entries/supererogation/>

Horgon, T., Timmons, M. (2010) 'Untying a Knot from the Inside Out: Reflections on the Paradox of Supererogation', in *Social Philosophy and Policy*, 27 (2) pp.29-63

Holton, R. (2004) 'Rational Resolve' in *Philosophical Review,* 113 (4) pp.507-535

Homiak, M. (2016) 'Moral Character' *The Stanford Encyclopedia of Philosophy (Fall 2016 Edition)* Edward N. Zalta (ed.). Available online at:
<https://plato.stanford.edu/archives/fall2016/entries/moral-character/>

Hume, D. (1985) *A Treatise of Human Nature*, London: Penguin Classics

Hurka, T. (2014) *British Ethical Theorists from Sidgwick to Ewing*, Oxford: Oxford University Press

Johnson, R. (2014) 'Kant's Moral Philosophy' *The Stanford Encyclopedia of Philosophy (Summer 2014 Edition)* Edward N. Zalta (ed.). Available online at:
<http://plato.stanford.edu/archives/sum2014/entries/kant-moral/>

Johnson, R., Cureton, A. (2018) 'Kant's Moral Philosophy' *The Stanford Encyclopedia of Philosophy (Spring 2018 Edition)* Edward N. Zalta (ed.). Available online at:
<https://plato.stanford.edu/archives/spr2018/entries/kant-moral/>

Joyce, R. (2001) *The Myth of Morality*, Cambridge: Cambridge University Press

Joyce, R. (2006) *The Evolution of Morality*, Cambridge, MA: MIT Press

Kagan, S. (1989) *The Limits of Morality*, Oxford: Clarendon Press

Kant, I. (2012) *Groundwork for the Metaphysics of Morals*, Cambridge: Cambridge University Press

Katz, L. (2016) "Pleasure", *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition),* Edward N. Zalta (ed.), Available online at:
<http://plato.stanford.edu/archives/win2016/entries/pleasure/>.

Kiesewetter, B. (2011) 'Ought and the Perspective of the Agent' in *Journal of Ethics and Social Philosophy*, 5 (3) pp.1-24

Kiesewetter, B. (2017) *The Normativity of Rationality*, Oxford: Oxford University Press

Kiesewetter, B. (forthcoming) 'What Kind of Perspectivism?' in *Journal of Moral Philosophy*

Kolodny, N. (2005) 'Why Be Rational?' in *Mind,* 114 pp.509-563

Korsgaard, C. (1993) 'The Reasons We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Values' in *Social Philosophy and Policy* 10 (1) pp.24-51

Korsgaard, C. (1996a) *The Sources of Normativity,* Cambridge: Cambridge University Press

Korsgaard, C. (1996b) *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press

Korsgaard, C. (2009) *Self-Constitution: Agency, Identity, and Integrity*, Oxford: Oxford University Press

Lang, G. (2012) 'What's the Matter? Review of Derek Parfit, On What Matters' in *Utilitas* 24 (2), pp.300-312

Littlejohn, C. (forthcoming) 'Being More Realistic About Reasons: On Rationality and Reasons Perspectivism' in *Philosophy and Phenomenological Research.*

Lord, E. (2015) 'Acting for the Right Reasons, Abilities and Obligation' in *Oxford Studies in Metaethics Volume 10,* Russ Shafer-Landau (ed.) Oxford: Oxford University Press

Mackie, J. (1977) *Ethics: Inventing Right and Wrong*, London: Penguin

Manne, K. (2013) 'On Being Social in Metaethics' in *Oxford Studies in Metaethics Volume 8* pp.50-74, Oxford: Oxford University Press

Manne, K. (2014) 'Internalism about Reasons: Sad But True?' in *Philosophical Studies* 167 no.1, pp.89-117

Manne, K. (forthcoming) 'Locating Morality: Moral Imperatives as Bodily Imperatives' in *Oxford Studies in Metaethics*

Markovits, J. (2014) *Moral Reasons,* Oxford: Oxford University Press

Mason, C. (2006) 'Internal Reasons and Practical Limits on Rational Deliberation' in *Philosophical Explorations,* 9 (2) pp.163-177

McDowell, J. (1995) 'Might There Be External Reasons?' in *World, Mind and Ethics*: *Essays on the Ethical Philosophy of Bernard Williams,* J. Altham and R. Harrison (eds.) Cambridge: Cambridge University Press

McElwee, B. (2007) 'Consequentialism, Demandingness and the Monism of Practical Reason' in *Proceedings of the Aristotelian* Society 107 pp.359-374

McElwee, B. (2016) 'Demandingness Objections in Ethics' in *The Philosophical Quarterly* 67 pp.84-105

McElwee, B. (2017) 'Supererogation Across Normative Domains' in *Australasian Journal of Philosophy,* 95 (3), pp.505-516

McLeod, O. (2001) 'Just Plain Ought' in *Journal of Ethics*, 5 (4), pp.269-291

Millgram, E. (1996) 'Williams' Argument Against External Reasons' in *Nous*, 30 (2), pp.197-220

Moore, A. (2013) 'Hedonism', in *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition) Edward N. Zalta (ed.), Available online at:
<https://plato.stanford.edu/archives/win2013/entries/hedonism/>

Murphy, D. (2015) "Philosophy of Psychiatry", in *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition) Edward N. Zalta (ed.), Available online at:
<https://plato.stanford.edu/archives/spr2015/entries/psychiatry/>.

Nagel, T. (1978) *The Possibility of Altruism*, Chichester: Princeton University Press

Norcross, A. (2006) 'The Scalar Approach to Utilitarianism' in West, H. (ed.) The *Blackwell Guide to Mill's Utilitarianism,* Oxford: Blackwell Publishing

Olson, J. (2014) *Moral Error Theory: History, Critique, Defence*, Oxford: Oxford University Press

Parfit, D. (1984) *Reasons and Persons*, Oxford: Clarendon Press

Parfit, D. (1997) 'Reasons and Motivation' in *Aristotelian Society Supplementary Volume* 71 (1) pp.99-130.

Parfit, D. (2011a) *On What Matters Vol.1*, Oxford: Oxford University Press

Parfit, D. (2011b) *On What Matters Vol.2,* Oxford: Oxford University Press

Pettit, P. (1993) 'Consequentialism' in P. Singer (ed.) *A Companian to Ethics*, Oxford: Blackwell Publishing Ltd.

Pettit, P., Smith, M. (1990) 'Backgrounding Desire' in *Philosophical Review*, vol.99 no.4 pp.565-592

Plato (2000) *Philebus*, Infomotions, Inc., South Bend, US. Available from: ProQuest ebrary. [15 November 2016].

Piller, C. (2013) 'The Bootstrapping Objection' in *Organon F: Medzinárodný Časopis Pre Analytickú Filozofiu,* 20 (4) pp.612-631

Rachels, S. (2000) 'Is Unpleasantness Intrinsic to Unpleasant Experiences?' in *Philosophical Studies* 99 (2) pp.187-210

Railton, P. (1984) 'Alienation, Consequentialism and the Demands of Morality' in *Philosophy & Public Affairs,* 13 (2) pp.134-171

Railton, P. (1997) 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action' in Garrett Cullity & Berys Nigel Gaut (eds.) *Ethics and Practical Reason*, Oxford: Oxford University Press.

Ratcliffe, M. (2015) *Experiences of Depression: a study in phenomenology,* Oxford: Oxford University Press

Rosen, G. (2004) 'Skepticism about Moral Responsibility' in *Philosophical Perspectives,* 18 (1) pp.295-313

Scanlon, T. (1998) *What We Owe To Each Other*, London: Harvard University Press

Scheffler, S. (1992) *Human Morality* Oxford: Oxford University Press

Scheffler, S. (1994) *The Rejection of Consequentialism*, Oxford: Clarendon Press

Schroeder, M. (2004) *Slaves of the Passions*, Oxford: Oxford University Press

Schroeder, M. (2005) 'The Hypothetical Imperative?' in *Australasian Journal of Philosophy* 83 (3) pp.357-372

Schroeder, T. (2017) "Desire", *Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), Available online at:
<http://plato.stanford.edu/archives/sum2015/entries/desire/>.

Shaver, R. (2016) 'Sidgwick on Pleasure' in *Ethics* 126, pp.901-928

Shepski, L. (2008) 'The Vanishing Argument from Queerness' in *Australasian Journal of Philosophy* 86 (3) pp.371-387

Sidgwick, H. (1907) *The Methods of Ethics*, London: Macmillan and Co.

Singer, P. (2009) 'Reply to Michael Huemer', in J. Schaler (ed.) *Peter Singer Under Fire,* Chicago and La Salle: Open Court Publishing Company

Smith, M. (1987) 'The Humean Theory of Motivation' in *Mind*, 96 (381) pp.36-61

Smith, M. (1994) *The Moral Problem,* Oxford: Blackwell Publishing

Smith, M. (2004) 'Instrumental Desires, Instrumental Rationality,' *Proceedings of the Aristotelian Society* (Supplementary Volume), 78 (1) pp.93–109

Smith, M., Lewis, D., Johnston, M. (1989) 'Dispositional Theories of Value' in *Proceedings of the Aristotelian Society* 63 pp.89-174

Skorupski, J. (2007b) 'Buck-Passing About Goodness' in J. Josefsson D. Egonsson (ed.), *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.*

Sobel, D. (2011) 'Parfit's Case Against Subjectivism' in *Oxford Studies in Metaethics Vol.6,* Oxford: Oxford University Press

Sobel, D., Copp, D. (2001) 'Against Direction of Fit Accounts of Belief and Desire' in *Analysis* 61: (1) pp.44-53.

Street, S. (2005) 'A Darwinian Dilemma for Realist Theories of Value' in *Philosophical Studies* 127 pp.109-166

Street, S. (2009) 'In Defence of Future Tuesday Indifference' in *Philosophical Issues* 19, pp.273-289

Streumer, B. (2007) 'Reasons and Impossibility' in *Philosophical Studies* 136 pp.351-384

Stroud, S. (1998) 'Moral Overridingness and Moral Theory' in *Pacific Philosophical Quarterly* 79

Thomson, J. (2008) *Normativity,* Chicago and La Salle: Open Court

Urmson, J. (1958) 'Saints and Heroes' in A. Melden (ed.) *Essays in Moral Philosophy*, Seattle: University of Washington Press

Wallace, R. J. (2006) *Normativity and the Will: Selected Essays on Moral Psychology and Practical Reason,* Oxford: Oxford University Press

Way, J. & Whiting, D. (2017) 'Perspectivism and the Argument from Guidance' in *Ethical Theory and Moral Practice,* 20 (2) pp.361-374

Wedgwood, R. (2011) 'Instrumental Rationality' in *Oxford Studies in Metaethics volume 6,* Russ Shafer-Landau (ed.) pp.280-309

Williams, B. (1981) 'Internal and External Reasons' in *Moral Luck*, Cambridge: Cambridge University Press

Williams, B. (1995) 'Internal Reasons and the Obscurity of Blame' in *Making Sense of Humanity,* Cambridge: Cambridge University Press

Williams, B. (2011) *Ethics and the Limits of Philosophy*, Abingdon: Routledge.

Wolf, S. (1982) 'Moral Saints' in *The Journal of Philosophy* 79 (9) pp.419-439

Woollard, F. (2010) Review: Peter Singer Under Fire*, Philosophical Reviews*
Available online at:
<http://ndpr.nd.edu/news/24252-peter-singer-under-fire-the-moral-iconoclast-faces-his-critics/>

Zimmerman, M. (2008) *Living With Uncertainty*, Cambridge: Cambridge University Press