Fitting Consequentialism

Richard Yetter Chappell

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Philosophy

Advisors: Michael Smith and Philip Pettit

September 2012

Abstract

According to the Fitting Attitudes analysis of value, we can understand *value* in terms of *desirability* or what it is fitting to desire. But we can also raise normative questions about the fittingness of (e.g.) beliefs, emotions, and choices. My dissertation explores the broader significance of such 'fittingness' evaluations from a consequentialist standpoint. This project has both a normative and a metaethical component. The normative component develops and assesses the consequentialist's conception of a morally fitting (or virtuous) agent, thereby responding to several traditional character-based objections to the view. Critics have alleged, for example, that a consequentialist agent would see people as mere value receptacles, be cold and calculating, untrustworthy, and disturbingly 'alienated' from themselves and others. I rebut these charges. The metaethical component of my project explores how attention to the possible types of 'fittingness' evaluations can help us to define the scope and limits of normative theorizing, with important implications for how we should formulate consequentialism. In particular, I argue that even if we think that what's right is *determined by* the value facts, we should not go so far as to think that rightness is conceptually *reducible* to goodness.

Acknowledgements

I would like to thank Michael Smith, Philip Pettit, Peter Singer, and Tim Mulgan for their very helpful comments and advice on multiple chapter drafts. I'm also grateful to the readers and commentators of my blog at `www.philosophyetc.net`, where many of the ideas in my dissertation first took form. But by far my greatest debt is to my wife Helen Yetter Chappell, who has aided me every step of the way, from brainstorming and working through rough ideas, to editing final drafts.

For helpful comments on particular chapters, I would also like to thank Ryan Cook, Nate Gadd, Hrishikesh Joshi, Rob Lawlor, Eden Lin, Errol Lord, Brennan McDavid, Sarah McGrath, Doug Portmore, Andrew Reisner, Derek Shiller, Nic Southwood, and audiences at the ANU, Bowling Green State University, and Princeton's Dissertation Seminar.

The first chapter, 'Fittingness: The Sole Normative Primitive', has been accepted for publication in *The Philosophical Quarterly*.

For Helen.

Contents

Introduction

This dissertation explores connections between *consequentialist* moral theory and *fittingness* evaluations.

Consequentialism is the view that what we ought to do is determined by how well things would (or could be expected to) turn out, on each of the options available to us. The simplest such view, Maximizing Act Consequentialism, specifies that we ought to choose the best of our available options. That is, we should choose whatever action would do the most good, or bring about the most desirable outcome. Utilitarianism further specifies that the value of an outcome is determined by the well-being of sentient creatures, impartially considered. On this view, we ought to do whatever will make people (broadly construed) best off on net.

Utilitarianism is structurally simple in that it offers us neither "options" nor "constraints". We might, for example, be morally required to kill ourselves if that would benefit others more than it harms us. We neither have the moral *option* to favour our own interests, and nor are we deontologically *constrained* from performing acts of distasteful types like *killing*, so long as the benefits achieved outweigh the harms imposed by so acting. At some points in this dissertation I will relax the assumption of strict impartiality, but throughout I will defend the core consequentialist idea that there are no deontological constraints on moral action, and that what ultimately matters is just to make the world as good as possible (whether from an impartial or partial perspective).

Fittingness evaluations concern whether an attitude is *merited* by its object. For example, beliefs are fitting when the thing believed is true (or supported by credible evidence), desires are fitting when their objects are truly good or desirable, and the emotion of fear is fitting when faced with a threat that is genuinely fearsome.

According to the Fitting Attitudes analysis of value, we can understand *value* in terms of *desirability* or what it is fitting to desire. But we can also raise normative questions about the fittingness of (e.g.) beliefs, emotions, and choices. This dissertation explores the broader significance of such 'fittingness' evaluations from a consequentialist standpoint. This project has both a normative and a metaethical component. The normative component develops and assesses the Consequentialist's conception of a morally fitting (or virtuous) agent, thereby responding to several traditional character-based objections to the view. Critics have alleged, for example, that a consequentialist agent would see people as mere value receptacles, be cold and calculating, untrustworthy, and disturbingly 'alienated' from themselves and others. I rebut these charges. The metaethical component of my project explores how attention to the possible objects of 'fittingness' evaluations can help us to define the scope and limits of normative theorizing, with important implications for how we should formulate consequentialism. In particular, I argue that even if we think that what's right is *determined by* the value facts, we should not go so far as to think that rightness is conceptually *reducible* to goodness.

Chapter 1. Fittingness: The Sole Normative Primitive

Consequentialists have traditionally followed G.E. Moore in taking the concept of *value* as their normative primitive. This metaethical view may be accompanied by the assumption that axiology exhausts normative ethics: i.e., that it's a substantive question what's valuable (good or bad), but there's no *further* question to ask about 'right' and 'wrong'. Global Consequentialists, in particular, have embraced the idea

that we should not *privilege* acts as uniquely subject to normative evaluation in terms of 'rightness'. Instead, they hold, any kind of normative assessment that we can make of acts, we can just as well make of any other possible object of evaluation — from eye colours to climates.

An alternative tradition follows Sidgwick in understanding value as a matter of *desirability*, or what we have (fitting) reason to desire.[1] In this chapter, I make the case for taking this notion of *fittingness* as our normative primitive. The Moorean's value claims can be captured within the fittingness framework as a matter of the object's being *fit to desire*. But by appealing to the notion of 'fittingness' more generally, we expand our theory's expressive power. Not only can we talk about value or what's fit to desire, but also what's fit to believe, or — more importantly for our purposes — what it is fitting (right) to *choose to do*. This conceptual framework allows us to make substantive new claims about the rightness of acts, that are not merely disguised claims about value, and for which there is no analogous normative status that applies to mere evaluands like eye colours. I thus argue against Global Consequentialism.

Chapter 2. The Fitting and the Fortunate

Critics of consequentialism often object to *how a consequentialist agent would (allegedly) think*. They claim that the consequentialist agent is, in some sense, a *bad* character: cold and calculating, incapable of genuine friendship, etc. Defenders of consequentialism typically dismiss such objections by citing the distinction between 'criteria of rightness' and 'decision procedures'. (Utility provides the criterion that determines the moral status of an act, but it's a further question whether agents ought to attempt to calculate utilities themselves. If it'd have bad results then consequentialists would recommend against it!) Although there's something right about

[1] Note that what it's fitting (or rational) to desire may come apart from what we actually desire.

this response, it is not entirely satisfactory. I argue that the strongest objection in this vicinity is not just that thinking like a consequentialist would have bad results, but that such a psychology would be morally *perverse*, in a sense that's incompatible with the psychology in question constituting a morally *accurate* way of thinking.

To begin, we must distinguish two very different kinds of normative evaluation: the 'fortunate', and the 'fitting'. If a psychological trait is *fortunate*, or instrumentally valuable, then it will be recommended by the normative theory as something to aim *at*. Roughly, this is to say that it is desirable, or ought to be inculcated, or such like. For example, if caring about your friends is conducive to your own happiness, then hedonistic egoism would evaluate such concerns positively in this sense. On the other hand, we can ask whether the agent's psychology *fits with* or *embodies* the normative theory — roughly, whether the agent is responsive to the reasons posited by the theory: whether he desires what the theory says is desirable, etc. This is to ask whether the agent's psychology is, in a sense, appropriate, rational, virtuous, or (as I will say) 'fitting'. These two kinds of normative evaluation may come apart, as in the so-called 'paradox of hedonism': one is more likely to achieve hedonistically-fortunate results (i.e., happiness) if one does not possess the fitting-hedonistic mindset of adopting happiness as one's supreme goal.

With this distinction in hand, we can now distinguish two interpretations of the "bad character" objection to consequentialism. One claims that the fitting consequentialist psychology is *unfortunate*, or that the theory is 'self-effacing'. But this is a poor objection: it's no fault of a moral theory if events conspire to punish those who believe and act on the truth, though it might well mean that it'd be better, from a moral perspective, if we believed some false moral theory instead. Instead, when deontologists complain about the bad character of a committed consequentialist agent, we should interpret them as leveling the stronger objection that the fitting consequentialist psychology is (contrary to the consequentialist's claims) *not actually*

morally fitting. For example, they argue that the consequentialist agent is incapable of friendship or commitment to projects — but, they may add, this seems like an intrinsic defect: surely *genuine* virtue and rationality are not incompatible with these important goods. So, they conclude, the consequentialist's conception of rationality (virtue, fittingness) must be in error.

This objection is the real challenge. Consequentialists have typically neglected it, because they have focused exclusively on evaluations of fortunateness. They haven't appreciated that their theory also commits them to a conception of the morally *fitting* agent. To take up this challenge, we must either (i) bite the bullet and insist that what the deontologist identifies as moral 'defects' are not really so, or else (ii) argue that, properly understood, the fitting consequentialist agent would not in fact possess the identified defects.

In this chapter, I begin the latter task of rehabilitating our conception of the fitting consequentialist agent. For example, some traditional objections to utilitarianism assume that the fitting utilitarian agent would have but a single desire: to maximize utility — an end to which any individual's welfare is a merely fungible means. I argue that this is a misconception. By attending more closely to the structure of their value theory, we can see that the utilitarian agent is instead committed to separately desiring the welfare of *each* individual for their own sake, and not merely as a means.

Chapter 3. What's Fit for the Fallible

In this chapter, I further pursue the challenge of developing a positive picture of the fitting consequentialist agent — with a particular focus on what's rationally fitting for fallible creatures with human-sized minds. In so doing, I defuse several traditional objections:

1. that the consequentialist would be constantly calculating (or, as Williams put it, have 'one thought too many' before acting to save a loved one);

2. that they would be emotionally narcissistic, actively "regret[ting] the absence or lack of any and every attainable good" (Stocker);

3. that they would be unstable and unpredictable, ready to break a promise at the slightest hint that doing so would be even *slightly* more conducive to the overall good.

I also clarify how this project relates to the 'sophisticated' or 'two-level' consequentialisms developed by Railton and R.M. Hare. Insofar as their 'good consequentialist agent' possesses whatever motivations would be most desirable, what they are describing is merely a *fortunate*, not necessarily any kind of *fitting*, psychology. We can imagine circumstances in which it would be most fortunate to possess positively malicious motivations, but that clearly wouldn't make malice a virtue (or 'fitting', as I use the term here). I suggest that we can identify a proper subset of desirable dispositions as *rationality-enhancing*, in contrast to others that merely cause better results by some external, non-rational means. It is only the rationality-enhancing dispositions that we should attribute to the fitting moral agent: externally desirable dispositions are worth pursuing, but once achieved may disqualify one's resulting psychology from being morally fitting. (Just as, for example, a hedonist should try to acquire non-instrumental interests of other kinds, but once they do they no longer qualify as fitting the theory of hedonism. They have rationally made themselves irrational, by hedonistic lights.)

This distinction can help illuminate the Gauthier-Parfit debate about rational transmission: i.e., whether it's rational to act on a disposition that it's rational to acquire. I argue that Parfit is correct that no such general principle holds: even if the deterrence value renders it rational to acquire a doomsday disposition, it is

not necessarily rational to *act* on such a disposition. Nonetheless, we may hold out hope for a related version of the transmission principle that is restricted to those dispositions that are independently identifiable as 'rationality-enhancing'.

Chapter 4. Virtue and Salience

This chapter further develops 'non-ideal' rational (fittingness) norms for beings with limited attention. I argue that alleged 'virtues of ignorance' (e.g. modesty, believing better of friends than the evidence supports, etc.) are better understood as 'virtues of salience'. The modest person, for example, needn't have any *false beliefs* about their own accomplishments; what sets them apart from the immodest is instead that their own accomplishments aren't as *salient* in their thoughts — their attention is not constantly directed back towards themselves in the manner of the immodest.

I also explore the role of salience in ordinary evaluations of the demands of beneficence. Ordinary intuitions support the idea that failing to give to famine relief (for example) is in some sense *less bad* than failing to help someone who is right before your eyes. While it cannot plausibly be contended that saving nearby children is more important in principle (as if entering your visual field somehow increased their moral status!), ignoring *salient* needs reveals a greater deficit of benevolent motivation in the agent, and hence renders them more blameworthy. We can thus go some way towards reconciling Consequentialism's extreme demands with commonsense moral distinctions.

Chapter 1

Fittingness: The Sole Normative Primitive.

This chapter draws on the 'Fitting Attitudes' analysis of value to argue that we should take the concept of *fittingness* (rather than *value*) as our normative primitive. I will argue that the fittingness framework enhances the clarity and expressive power of our normative theorizing. Along the way, we will see how the fittingness framework illuminates our understanding of various moral theories, and why it casts doubt on the Global Consequentialist idea that acts and (say) eye colours are normatively on a par. We will see why even consequentialists, in taking rightness to be in some sense *determined by* goodness, should not think that rightness is conceptually *reducible* to goodness. Finally, I will use the fittingness view to explicate the distinction between consequentialist and deontological theories, with particular attention to the contentious case of Rule Consequentialism.

1.1 Introduction

Consequentialists have traditionally followed Moore (1903a) in taking the concept of *value* as their normative primitive. This metaethical view may be accompanied

by the assumption that axiology exhausts normative ethics: i.e., it's a substantive question what's valuable (good or bad), but there's no *further* question of 'right' and 'wrong'. Some recent theorists have embraced the implication that we should not privilege acts as uniquely subject to normative evaluation in terms of 'rightness'. Instead, they claim, any normative status that's applicable to acts, can just as well apply to other possible objects of evaluation, from eye colours to climates. Call this the **Parity Thesis**. Such parity might be achieved by either shrinking or extending the domain of 'rightness'. These two possibilities are reflected in Scalar and Global Consequentialisms, respectively. Global Consequentialism defines the 'right' x (for any category of evaluand x, be it acts, climates, or whatever) as the best x of those available (Pettit and Smith 2000). It thus indiscriminately extends 'rightness' talk to all evaluands. Scalar Consequentialism, on the other hand, does away with 'rightness' altogether, and merely has us evaluate acts (like anything else) as more or less good.[1]

An alternative tradition follows Brentano (1902) and Sidgwick (1907)[2] in understanding value in terms of a more fundamental notion of *desirability* or what we have (fitting) reason to desire.[3] In this chapter, I make the case for taking this notion of *fittingness* as the fundamental concept for use in normative theorizing. The Moorean's

[1] At least, this is one natural form that Scalar Consequentialism might take. Call it *Evaluative Scalar Consequentialism*: the view that *all* we can say about actions is that they have more or less value, and there's no other normative claims — not even about, say, reasons for action — to make at all. (I see this as a terminological variant of Global Consequentialism, which will be my main foil for the remainder of the chapter.) On the other hand, we can imagine a version of the view which, despite rejecting the *binary* distinction between 'right' and 'wrong', nonetheless allows that we can indeed make normative claims about choiceworthiness (or reasons for action) over and above evaluative claims, it's just that this further normative status is also a matter of degree. This more sophisticated view — call it *Normative Scalar Consequentialism* — is compatible with my rejection of the parity thesis.

I'm not certain which version Norcross is advocating in his (2006), for while he does speak of 'reasons for action', it isn't clear whether or not he thinks such claims are conceptually reducible to value claims.

[2] Andrew Reisner pointed out to me that it's controversial to what extent we should characterize Sidgwick as a fitting attitudes theorist — Moore himself did not think of Sidgwick as such (Moore 1903b). Nonetheless, I follow Smith (2010) in finding Sidgwick's talk of 'desirability' and 'rational ends', e.g. in section I.ix.3 of his (1907), at least highly suggestive.

[3] Scanlon (1998) and other deontologists understand value more broadly as warranting various kinds of pro-attitude (e.g. admiration, respect). For simplicity, I will focus exclusively on the consequentialist conception of value as desirability.

value claims can still be captured within the fittingness framework as a matter of the object's being *fit to desire*. But by appealing to the notion of 'fittingness' more generally, we expand our theory's expressive power. Not only can we talk about value (what's fit to desire), but also what's fit to believe, and — more importantly for our purposes — what it is fitting (right)[4] to *choose to do*. This conceptual framework allows us to make substantive new claims about the rightness of acts, that are not merely disguised claims about value restated in different words. We will also see that the fittingness view provides a more ecumenical framework for normative theorizing, allowing logical space for both consequentialist and deontological theories to emerge.

1.1.1 Conceptual Frameworks

To get a firmer grip on the notion, we may understand a 'conceptual framework' to be a set of interrelated concepts, with some specified as primitives, and others defined in terms of the primitives. The conceptual framework for a domain is, in effect, the expressive toolkit through which we can make first order claims on some topic. In the case of the normative domain, I suggest three desiderata against which we may assess candidate conceptual frameworks:

1. It must provide adequate conceptual resources for us to *express* any expressible normative truth.[5]

[4] Admittedly, the notion of fitting choice does not exactly correspond to the ordinary deontic notions of obligation or permissibility. In particular, as noted in fn.1, I don't mean to rule out the kind of 'normative scalar consequentialism' that treats moral rightness as coming in degrees, with no sharp line between those acts that are sufficiently fitting or justified to qualify as 'permissible' and those that aren't. For current purposes, the relevant similarity to our ordinary talk of 'right action' is just that we're talking about a kind of normative status that applies distinctively to one's choice of action. For more on the fuzziness of 'rightness'-talk, see Lawlor (2009, Appendix A).

[5] Of course, most normative claims involve the use of non-normative concepts too. So a more precise statement of the first desideratum is that the conceptual framework for our normative domain must provide adequate conceptual resources for expressing the *normative parts* of any expressible normative truth. It may require supplementation by non-normative conceptual frameworks in order to express the non-normative parts of a normative claim. Hopefully the intended idea is clear enough.

2. It should, so far as possible, remain neutral on the disputes of first-order normative ethics.

3. It should reflect and illuminate any natural '*joints*' in the ontological structure of the normative domain (insofar as the normative domain is structured).

These desiderata are motivated by the philosophical role that we are seeking to fill. We are looking for a conceptual framework within which to conduct our first-order normative inquiry. A candidate conceptual framework will be of little use to us if we cannot use it to express the answers we seek. Gross non-neutrality in a framework similarly precludes using it for this purpose, insofar as it simply presupposes answers to the questions we wished to ask. The third desideratum is less strictly essential, but clearly a nice feature to have if possible. For example, it may help illuminate which first-order debates are genuinely significant, whereas a framework with (e.g.) redundant concepts risks encouraging terminological disputes and other confusions.

1.1.2 Roadmap

The structure of this chapter is as follows. In section 1.2, I motivate the fittingness view in relation to other positions in the literature, with particular reference to its expressive power. Section 1.3 pits the fittingness framework against value primitivism with respect to the third desideratum. To anticipate: the fittingness view implies that there is a distinct normative status corresponding to each kind of judgment-sensitive 'rational output' that can be assessed as fitting or unfitting to the situation, e.g., beliefs, desires, and actions. So, in particular, there's a distinctive normative status that is applicable to actions, with no analogue for mere evaluands like eye colours. I offer several reasons to think that this accurately reflects an important 'structural' difference between judgment-sensitive and non-judgment-sensitive evaluands, thus arguing against the value primitivist's Parity Thesis, and by extension, against Global Consequentialism.

Section 1.4 uses the above asymmetry in the fittingness framework to defend Act Consequentialism against the Global Consequentialist's charge that, in rejecting the Parity Thesis, we are "privileging" acts in a theoretically objectionable way. Whereas advocates of Global Consequentialism have claimed that theirs is the view that Act Consequentialists were striving for all along, I argue that the reverse is true: Global Consequentialists who accept the fittingness view should be led to Act Consequentialism.

Section 1.5 defuses an important objection by showing how to understand the structure of deontological and rule consequentialist theories in terms of the fittingness view. Finally, in section 1.6, I consider the objection that facts about *fitting choice* may reduce to facts about *fitting desire*. I argue that, while the two may typically coincide, no such reduction is possible, at least not without collapsing the consequentialism-deontology distinction.

1.2 Reasons-Talk and Fitting Attitudes

To begin, let me clarify how Scanlonian 'reasons fundamentalism' (the view that takes *reasons* as the fundamental normative concept) fits into the dispute between the fittingness view and value primitivism. I see the fittingness framework as very much in the spirit of reasons fundamentalism. My worry about starting with reasons-talk is that in practice it seems to invite unnecessary ambiguity and confusion. Normative reasons are typically introduced as 'facts that count in favour' of an act or attitude, but there are two very different kinds of 'favouring' that might be invoked here.[6] This is seen, for example, in the distinction between 'epistemic' and 'practical' reasons for belief. Suppose that a prankster demon will reward me for believing that grass is

[6] This distinction is typically analysed in terms of 'object-given' versus 'state-given' reasons for attitudes. I avoid this terminology because there may be tricky cases where the object of the attitude is in some sense responsible for our having what should intuitively be characterized as a state-given reason. Cf. Rabinowicz and Rønnow-Rasmussen (2004).

purple. This fact might be said to 'count in favour' of the belief, in one obvious sense: it establishes that there is (instrumental) value to my having this belief — it would be a *fortunate* state for me to be in. But there's a very different sense in which evidence, e.g. the fact *that grass looks green*, 'counts against' believing grass to be purple. Evidence counts for or against belief in a proposition by speaking to whether the proposition is true and hence belief-worthy or *fit* to believe.

We can similarly distinguish value-based reasons and fitting reasons for *desire*, say if we would be rewarded for desiring that others suffer. Suffering is undesirable: it has qualities that 'count against' desiring it, in the sense of rendering such desires *unfitting* to their objects. But external incentives may 'count in favour' of such perverse desires, by rendering their possession instrumentally valuable.

In light of this ambiguity, it is unhelpful to take the notion of 'counting in favour' as primitive (in our normative theorizing). We must further specify that our conception of 'reasons' is that of counting in favour *in the sense of* rendering an attitude fitting, rather than merely fortunate. Otherwise, the inclusion of value-based reasons leads to the 'Wrong Kind of Reasons' problem (Rabinowicz and Rønnow-Rasmussen 2004) for analysing value in terms of reasons for desire. That is, an external incentive to desire suffering would then suffice to give you "reason to desire" suffering, though it clearly wouldn't make suffering itself good, thus undermining the identification of what's good with what we have reason to desire.

We avoid this problem if we start from the unambiguous notion of fittingness: external incentives to desire suffering don't make suffering itself any more desirable (fit to desire), after all. So rather than taking reasons directly as primitive, we do better to reconstruct the notion of 'reasons for desire' explicitly in terms of *desirability characteristics* — features in virtue of which an object is fit to desire.[7] This allows

[7] Rather than starting from the *features* that render an object fitting to desire, we might follow Smith (1994, 2010) in taking as fundamental the *ideally rational psychology*, and then understand reasons for desire (or desirability characteristics) as *features that would figure in the ideal agent's desires*. I mean to remain neutral on this.

us to better avoid objections when it comes time to analyse value in terms of our normative primitive.

The unhelpful ambiguity likewise infects other general normative terms like 'ought'. There seems a sense in which you ought to believe/desire what's fitting, and another sense in which we can say that you 'ought' to believe/desire in whatever way would be best. It doesn't seem like there's any way to balance and combine these into a single normative conclusion; rather, the two kinds of evaluations are giving different answers to different questions.[8] So rather than starting with general normative terms like 'reasons', 'counts in favour', 'ought', or the like, we should start from our understanding of the two more specific kinds of normative evaluation: the *fitting* (or rationally warranted)[9] and the *fortunate* (or instrumentally valuable).

We need to be able to make both kinds of evaluation, in order to make sense of the cases discussed above. But we can achieve this using fittingness evaluations alone. For whenever a mental state would be fortunate, that means that it would be desirable — fitting to desire — to possess that mental state. To illustrate: the *belief* that grass is purple is unfitting,[10] but it's perfectly reasonable to *desire* to have the belief, in

[8] As we'll see below, we can understand the two questions as (1) is p believable/desirable? and (2) is the belief/desire that p desirable? Even if the answers diverge (suppose the answers are 'no' and 'yes', respectively), an agent could be fit in both respects: that is, he could reasonably fail to believe p, whilst reasonably desiring to possess the irrational-but-fortunate belief. Cf. Parfit (2011, Appendix A).

[9] I don't here mean to take any stand on the issue of 'objective' vs. 'subjective' oughts, or how incomplete and misleading evidence affects the normative status of attitudes. I'm inclined towards a pluralistic stance: there's an 'objective' sense of fittingness that corresponds to objective reasons, or what we might call 'objective rationality' — roughly, what would be rational given full information, modulo the conditional fallacy (Shope 1978). I take there to also be various 'subjective' or 'evidence-relative' senses of fittingness, which correspond to ordinary evaluations of rationality or reasonableness for fallible non-omniscient agents. For simplicity, we may focus on fully-informed ideal agents, for whom 'objective' and 'subjective' norms coincide.

[10] Some people report finding it just as intuitive, to their ear, to apply the word 'fitting' to instrumentally valuable beliefs like this. I mean to use 'fittingness' as a technical term to pick out precisely that intuitive normative concept that assesses (e.g.) desires according to whether their objects are genuinely desirable, and beliefs according to whether their object merits belief, etc. In other words, it is the normative concept which encompasses 'epistemic' rather than 'practical' reasons for belief, generalized to apply to other attitudes as well. You should associate it with words like 'warranted', 'appropriate', 'right' or 'correct', and (on the negative side) 'crazy' and 'perverse'; **not** with 'useful', 'fortunate', 'important', or (on the negative side) 'disastrous'. I hope readers find it intuitively compelling that there is a general notion in this vicinity which is the common thread between credible

light of the proffered reward. Likewise, though others' suffering is itself undesirable, it may be fitting to desire (for instrumental reasons) *that you (perversely) desire that others suffer*. In this way, value claims are analysable in terms of fittingness claims, allowing us to take the latter as our sole normative primitive without any cost to the expressive power of our normative theorising.

One might wonder whether we could just as well reverse this approach, and seek an analysis of fittingness in terms of value. But the prospects for such an analysis seem dim. For while there is an intimate conceptual connection between value and desirability, there is not any such obvious connection between value and (say) believability. The evidence may render belief in P fitting even when both P itself and *your believing of* P are unfortunate or disvaluable. We can even imagine worlds governed by truth-hating demons, where it is not even good *as a general rule* to believe what's supported by the evidence. So there seems little hope for analysing fitting attitudes in terms of value, the way that we can easily analyse value in terms of fitting attitudes.

This asymmetry suggests that the Fittingness framework has greater expressive power in the following sense: Any value claim can also be expressed in terms of fittingness, whereas it is not the case that all fittingness claims can also be expressed in terms of value. If we want to be able to make both kinds of claims, using only one kind of primitive, then this suggests the primitive we should choose is that of *fittingness*, rather than *value*.

My remarks so far have suggested a strong presumptive case for taking *fittingness* as our normative primitive. This is a variation of a view held by many philosophers, including Scanlon (1998, 2009) and Darwall (2003). The original contribution of this chapter comes next, as I develop the framework and demonstrate its theoretical payoff.

---

beliefs, desiring what's desirable, and fearing the fearsome. There is likewise something commonly *inappropriate* about crazy/incredible beliefs, perverse desires, and irrational phobias. We may capture this commonality by saying that each attitude is, in its own way, an *unfitting* response to its object.

1.3 Rational Outputs and the Scope of Normative Theorizing

The scope of our normative theorizing is constrained by our normative primitives. Everything there is to say can be said using the primitive concepts — otherwise we'd need additional primitives. So, if Mooreans are right to take the concept of *value* as their sole normative primitive, it must be that the value facts in some sense exhaust the normative facts: there's nothing more to say — or, at least, nothing distinctively *normative* — once we've settled what's good and bad. In particular, this means that there is no normative status of 'rightness' or fittingness that applies distinctively to acts or choices, as opposed to mere evaluands like eye colours and the global climate. All we can say is that some acts are better or worse than others (in virtue of leading to better or worse outcomes), in just the same way that some eye colours may be better or worse than others (in virtue of leading to better or worse outcomes). If we want to apply the term 'right' to the instrumentally best acts, we may as well call the instrumentally best eye colour the 'right' one to have too — though of course this is not to make a substantive new claim, but just to re-state the old evaluative claim using new words. In this way, value primitivists are naturally led to a kind of Global Consequentialism (Pettit and Smith 2000; Kagan 2000; Ord 2009), which we may examine here as a kind of case study of the Parity Thesis.

The essence of Global Consequentialism, as it interests me here,[11] is its symmetrical treatment of acts and other, non-judgment-sensitive objects of evaluation (what I call 'mere evaluands'). Pettit and Smith (2000, 122) write: "The crucial feature of global consequentialism is that it does not privilege any category of evaluand." Ord (ms) adds: "By encompassing all evaluands, rather than just the most prominent ones, it maximizes its expressive power while remaining nonarbitrary. Indeed, some

[11] One might offer a strictly weaker view that merely defends a global definition of *rightness*, whilst countenancing other normative evaluations — such as fittingness — that have non-global scope. But this would amount to a mere terminological variant of the act consequentialism that I defend in §1.4.

systems of global consequentialism promise to be simpler than act-consequentialism itself, for by allowing everything to be morally assessed [in the same way], they no longer need an associated theory of acts." (p.4) What's distinctive about this view is its whole-hearted embrace of the **Parity Thesis** that the normative domain is symmetrically structured, such that actions and eye colours are on a par when it comes to the ways in which we can normatively assess them. There's nothing special about judgment-sensitive evaluands, on this view. To think otherwise, Global Consequentialists claim, would be to arbitrarily privilege acts.

The value primitivist might try to escape this commitment to the parity thesis by constructing a new concept of 'rightness' out of a combination of the normative concept *value* plus the non-normative concept *can* (as it figures in the slogan, 'ought implies can').[12] So, for example, to call something 'right' might just be to say that: (i) the agent in question *could choose to do it*, and (ii) this choice would produce at least as much value as any other choice the agent could have made. Assuming that actions are just those things that an agent 'can' — in the relevant sense — do, it seems that the value primitivist has succeeded in securing access to a new normative term that applies to acts but not to eye colours, as desired.

I have two main objections to this. Firstly, it is not at all clear that this newly constructed concept really allows the value primitivist to make claims with any new *normative* impact. It's true that they can make a new claim that is partly normative, e.g. 'ϕ-ing is the right thing to do.' But this is decomposable into the conjunction of two claims: (i) the agent can ϕ, and (ii) no other option produces more value than ϕ-ing. Note that only the first claim is new (given a context in which we've already settled the value facts), whereas only the second claim is normative. So this move does not really allow the value primitivist to make additional normative claims in

[12] I owe this suggestion to Nic Southwood.

any interesting sense. It merely allows them the cheap victory of making conjunctive claims that are new in one respect and normative in another.

My second worry about this proposal is that it seems *ad hoc*, or lacking a principled basis for being restricted in scope to actions. After all, the normative part of the new conjunctive claim — i.e., the conjunct which claims that the object maximizes value — is just as applicable to eye colours as it is to actions. Presumably the underlying idea was meant to be that the conjunction of 'value' and 'can' claims is normatively significant in a way that goes beyond the mere sum of its parts (so to speak). This is a plausible claim, but I do not think it is one that the value primitivist can make. If we have an *independent* grasp of rightness, then we can claim that *acts that are the best an agent can do possess this further normative property of rightness*. But such a claim would be trivialized by the value primitivist's conceptual reduction of 'rightness' to 'the best an agent can do'. It would be what Parfit (2011, 71) calls a 'concealed tautology': equivalent to the claim that *acts that are the best an agent can do possess the property of being the best that agent can do*. This is not yet to say that there's anything normatively significant about being the best an agent can do, over and above the evaluative component of the claim. So I don't think that value primitivists can escape commitment to the parity thesis so easily.

Even so, the value primitivist might at least cushion the blow by pointing out some pragmatic reasons why we might tend to be more *interested* in evaluating actions than, say, eye colours.[13] Actions are the means by which we may voluntarily affect the world (and the amount of value in it). Moreover, our future behaviour may itself be influenced by praise and blame, whereas eye colours tend not to be so responsive. So even upon accepting the parity thesis, the value primitivist needn't face any problems explaining the act-centric focus of our ordinary moral practices. I'm happy to grant that. My objections to the parity thesis will emerge shortly (and more strongly

[13] Thanks to Philip Pettit, Andrew Reisner, and Peter Singer for pressing me on this point.

in section 1.6, where I argue that we need an independent concept of rightness in order to maintain the consequentialism-deontology distinction). For now, I just want to re-iterate that value primitivists are committed to seeing no deep or principled difference in the range of normative assessments we can make of acts and eye colours, respectively.

The Fittingness view provides us with more options. As we have seen, value claims can be understood as claims about what it is fitting to desire. Now, it's true that *this* sort of normative assessment applies without restriction: we can ask whether some eye colour is fit to be preferred over others, just as we can ask whether some choice or action is fit to be preferred over others. But that isn't all that there is to say about what's fitting. In addition to these questions of desirability, we can also ask what it is fitting to believe, feel, and *choose*.

This last option gives rise to a form of normative assessment that applies distinctively to acts (understood as the direct implementation of a choice), as a matter of principle. That is, in addition to asking whether some chosen act is fit *to desire*, we can ask whether it is *itself* fitting (qua choice).[14] No such additional question arises for eye colours. We can ask whether some eye colour is fit to desire, but there's no further question whether the eye colour is itself rationally fitting — eye colours just aren't subject to direct rational criticism in this way. It cannot be a rational failing for one's eyes to *be* a certain colour, the way it might be a rational failing (in some circumstances) to *choose to bring about* such an outcome.[15] The Fittingness

[14] To clarify the distinction between choice and desire: I take desire to be a persisting mental state that has some associated degree of weight. Any one desire, insofar as it may be outweighed by others, leaves open the question of what to do. Choice, on the other hand, is the process of *settling what to do*. In the case of Buridan's ass, caught between two indiscernible bales of hay, I take it that the fitting response is to desire *that you have either*, but to choose either particular one. See section 1.6 for further discussion.

[15] One may question whether acts are really subject to direct rational criticism, given that their rational status would seem to derive from the rationality of the *choice* to so act. I'm happy to grant that choice is the more fundamental object of assessment. But insofar as it is natural to understand the choice as partly constituting the action, an assessment of the choice is *ipso facto* an assessment of the action. Choices and eye colours, by contrast, stand in no such intimate relation.

view thus expands the scope of our moral theorizing to include a normative status distinctive to acts: whether they are fit to choose, or — in more colloquial terms — morally *right*.

I now want to argue that we should reject the Global Consequentialist's parity thesis, and hence favour the fittingness view over value primitivism. Consider again the completely general Global Consequentialist formula that *the 'right' x is simply the best (most utility-maximizing) x of those available*, where 'x' might be replaced by 'act', 'eye colour', or any other category of evaluand (Pettit and Smith 2000). I've been arguing that the Fittingness view shows there to be an important difference in normative structure here: there are some kinds of normative assessment that can be made of acts that cannot be made of eye colours. Though one may stipulatively define a sense of 'right' (or 'ought') that applies to everything including eye colours, there is an importantly natural, non-stipulative sense that applies only to judgment-sensitive evaluands (with no corresponding normative concept that applies just to non-judgment-sensitive evaluands). But this analysis will not convince someone who is not inclined to accept the fittingness view to begin with.

A more neutral way to bring out the difference in normative *structure* between actions and *mere* evaluands (like eye colour) is to think about how the relevant set of 'available options' is fixed in the above formula. In case of actions, the 'ought implies can' principle invokes a very particular sense of possibility: something like *rational power*, perhaps. That is, the relevant options are naturally restricted to something like *what the agent would succeed in doing if they had the intention to do it*[16] — though the precise details don't matter for my purposes.

The assessment of eye colours presents us with two notable points of contrast. First, there does not seem to be any naturally privileged set of options for mere evaluands; the relevant sense of 'availability' will instead presumably vary depending

[16] Thanks to Michael Smith for this suggestion.

on conversational context. (For example, one speaker may be interested in the eye colour that one 'ought' to have amongst those genetically possible for a child of one's parents, whereas another may be concerned to evaluate amongst all of those humanly possible. There's no fact of the matter as to which of these is the eye colour that you really 'ought' to have.) This looks an awful lot like ordinary evaluation — unlike in the case of actions, where the shift to 'ought'-talk brought with it a principled determination of the available options, and hence an unequivocal answer to the question of what act you ought to perform.

Secondly, eye colours are not judgment-sensitive, or directly subject to rational influence, the way that actions are. Judging that one ought to have blue eyes will not even tend to make it so. In this sense one cannot possess one's eye colour *for* (normative) reasons, the way that one can believe, desire, and act for reasons. We may demand that others justify their beliefs, or their actions, with adequate reasons. It does not make such sense to hold people (directly) accountable or answerable for their eye colours.

So while one might insist that there is some stipulative sense in which we "have reasons" to have the best available eye colour,[17] this verbal victory for the global consequentialist misses the point that there is a real difference in normative structure here that needs to be acknowledged. While we can evaluate anything, against any arbitrary set of 'alternatives', there is a special kind of normative assessment of rationally available actions that has no analogue for other (mere) evaluands. This difference allows us to make substantive claims about how we ought to act that are not just disguised evaluative claims. The same cannot be said for 'oughts' of eye-colour possession.

These facts suggest that in order to track the natural 'structure' of the normative domain — as per our third desideratum from §1.1.1 — normative theorists need a

[17] Thanks to Toby Ord and Michael Smith for pressing me on this point.

conceptual framework that offers asymmetric treatments of judgment-sensitive and non-judgment-sensitive evaluands. The Fittingness view achieves this by positing norms of 'fittingness' that apply directly to judgment-sensitive evaluands (belief, desire, action), one instance of which is the more specific notion of *fittingness to desire*, which in turn applies globally to all evaluands without restriction. The Fittingness view thus represents the normative domain as appropriately (asymmetrically) structured: our normative primitive of *fittingness* applies immediately to judgment-sensitive evaluands, and only indirectly (as mediated through the fittingness norms *for desire*) to everything else.

One needn't accept the Fittingness view in order to secure this desideratum, of course — a pluralist might combine the Moorean's value primitivism with an additional primitive concept of 'rightness'. But the Fittingness view is both more parsimonious and more powerful. We have seen that it is more parsimonious in unifying both evaluative and non-evaluative normative claims under the umbrella of the single normative primitive of 'fittingness'. What's more, it is more powerful in that it provides us with an explanation of *why* there is a distinctive form of normative assessment that applies only to acts. The explanation is that there are distinct forms of normative (fittingness) assessments corresponding to *each* of our judgment-sensitive rational outputs: e.g., beliefs, desires, and actions.

The Fittingness view tells us that the scope of normative theorizing is fixed by the range of our rational outputs, as those are the things that can be assessed as more or less fitting.[18] At a minimum, we can identify assessments of *fitting belief* as the province of epistemologists, *fitting desire* as the concern of axiologists (value theorists), and — I suggest — *fitting choice/action* as a further issue for normative ethicists. This list should hopefully provide a plausible enough starting point, but I should stress that it is ultimately up to our best moral psychology / theory of agency

[18] Darwall (2003) offers a similar suggestion: "In principle, there are as many normative notions as items (action and attitudes) that can be normatively regulated."

to identify any additional rational outputs (e.g. emotions), and hence to further delineate the scope for normative inquiry.

1.4 'Privileging' Acts

Act Consequentialism is the view that the rightness of an action is determined by the value of the resulting state of affairs (compared to the available alternatives). We may further specify that, according to act consequentialists, acts are the *only* items to which this normative status of 'rightness' may apply.[19] There is, in this sense, an important normative disparity claimed between acts and (say) eye colours. We are now in a position to put to work the theoretical apparatus developed in previous sections, illuminating the debate between Act and Global Consequentialists. In particular, we shall assess the charge that act consequentialism "privileges" actions in a theoretically objectionable way.

Let us first observe that *all* consequentialists start with an axiology, or theory of value, which tells us what things are valuable or fitting to desire. Given this starting point, the work that remains for their normative theory is to use these evaluative facts to derive some further normative claims about what we ought to do. We've seen that value primitivists are naturally led to the view that there isn't really anything more to say once we settle the value facts. All the substantive work is done by their axiology, and so to call an act 'right' is not to make a further claim at all, but is instead analytically equivalent to calling it 'best' or value-maximizing. The substantive content of this 'Global Consequentialism', over and above the axiology that might be shared by any other form of consequentialism, is purely negative:

[19] Technically, there may be an analogous status of 'rightness' or fittingness for other judgment-sensitive evaluands such as belief, etc. But epistemological questions and the like are outside the province of the moral theorist. The crucial point for now is just that the Act Consequentialist is naturally taken to reject the parity thesis, and to instead see an important asymmetry at least between acts and non-judgment-sensitive evaluands.

it's just the claim that there isn't anything more to say, beyond evaluating things (including acts) for their desirability.

But we've seen that there *is* more for a moral theory to say than just this. According to the fittingness view, in particular, we can assess acts not just for their desirability (whether they are fit *to desire*), but also for whether they *themselves* constitute fitting choices. This naturally supports Act Consequentialism over Global Consequentialism (modulo the qualms about 'rightness'-talk mentioned in footnote 4). Act Consequentialists begin with an axiology (an account of fitting desire), and to this they add just one new type of claim — a claim about what it is fitting (right) to *do*. The claim "an act is right if and only if (and because) it produces at least as much value as any available alternative" is no longer a mere analytic truth or stipulative definition, but a substantive normative claim that relates one normative property (fitting choice) to another (fitting desire).

Act Consequentialism invokes a normative status that applies distinctively to acts, and thus succeeds in going beyond its axiology to make substantive new claims, when Global Consequentialism does not. This suggests two reasons to prefer Act Consequentialism: first, because it has more substance, and secondly, because it achieves this by better respecting the structure and scope for normative theorizing revealed by the fittingness view and our best moral psychology. When Global Consequentialists deny that acts are subject to a distinctive form of normative evaluation (beyond the generally available assessment of global desirability), this is both factually incorrect and unnecessarily limiting.

I've argued that Global Consequentialists have erroneously neglected the structural differences between normatively-assessable acts and mere evaluands. It's a mistake for a normative theory to treat everything on a par, because it turns out that there's a distinctive normative question that arises for acts — namely, 'are they morally right, or fitting to choose?' — that has no analogue in case of mere eval-

uands. But let me add a conciliatory note. Global Consequentialists are right to insist that our *axiological* theory should assess each particular evaluand (including acts considered as mere evaluands) directly and in its own right. Many philosophers have made the contrary mistake of assuming that the best dispositions, for example, are simply those that result in the best actions.[20] To privilege actions in *this* (axiological) sense is indeed a straightforward error. It is important not to overlook the fact that dispositions of character can have good or bad effects other than via their manifestation in action (Adams 1976; Parfit 1984; Railton 1984; Pettit and Smith 2000). For example, it might be good, for the sake of deterrence, to have a transparent disposition to trigger a doomsday device when attacked. This does not entail that there's anything good about the act of mass destruction. The value here accrues from (transparently) possessing the disposition, not from exercising it.[21] So it would be a mistake to think that acts resulting from beneficial dispositions must themselves be beneficial. Each evaluand must be evaluated in its own right, as Global Consequentialists rightly remind us.

But just because actions should not be *axiologically* privileged, this does not mean that there are no important normative-theoretical differences between acts and (say) eye colours. Act Consequentialists can — and should — take care to avoid the axiological oversight that global consequentialists warn against, whilst maintaining that there's an important sense in which we can assess acts, but not eye colours, as morally right or wrong (over and above their being good or bad).

[20] For example, Shaw (2006, 13) writes: "Consequentialists generally assess dispositions, behavioral patterns, and character traits in the same instrumental way: one determines which ones are good, and how good they are, by looking at the actions they lead to."

Even Kagan (2000) arguably suffers from this oversight. After introducing the idea of evaluating rules according to their propensity to produce more or less good actions, Kagan claims that this is tantamount to "claim[ing] that rules should be evaluated *directly* in terms of the goodness of their consequences." (p.149) But this only follows if the downstream actions are the only possible consequences of internalizing a rule, which we'll see is false.

[21] Geoff Brennan pointed out to me that this is still a case where the disposition's value derives from its effect on (others') *actions*. But we can just as well construct cases where, e.g., the mere possession of a certain disposition tends to make one happier, or to produce good effects in the world, in a way that is unmediated by any actions.

To summarize my argument thus far: The fittingness view implies that the list of rational outputs (states or activities that are responsive to our normative judgments, and can be assessed as more or less rationally 'fitting') determines the scope of our normative theorizing. There are facts about which features of an agent are rational outputs, and hence what kinds of reasons there can be (e.g. reasons for belief, desire, and action). These moral psychological facts are arguably 'prior' to the question of what particular normative theory correctly specifies the contents of these various kinds of reasons. So our normative theorizing should acknowledge and respect these limitations. Consequentialists presuppose an account of our reasons for desire (i.e., their axiology). They have no ambition to override epistemologists' claims about our reasons for belief, nor do they tend to make claims about the rationality of other attitudinal states like emotions. This would seem to leave reasons for action as the only remaining target for further theorizing. So it is by no means arbitrary for Act Consequentialists to add further normative claims (i.e. besides what is already contained in their axiology) *only* about how we ought to act. This is the only kind of claim that *can* plausibly be added, once we have settled all questions about what we have reason to desire.

1.5 Analysing Rule Consequentialism

My analysis so far has presupposed that our theory of value settles all questions about what we have reason to desire: i.e., we should always prefer what's best.[22] However, deontologists and rule consequentialists might seem to deny this. (Here I'll focus on the latter for simplicity.) Rule consequentialists hold that we ought to act in accordance with certain rules (the general acceptance of which would maximize value) even on occasions when so acting is not itself value-maximizing (Hooker 2000). This

[22] Note that our theory of value need not be impartial ('agent-neutral'). If we each have agent-relative reasons to prefer the welfare of our own children over that of strangers, for example, then different states of affairs will be desirable relative to different people.

seems difficult to make sense of if the value facts exhaustively specify our reasons for desire, since by calling one act value-maximizing (worth preferring) but then prescribing another as worth choosing, rule consequentialism would seem to imply that we ought *to hope* that we act differently from how we ought *to act*. Such a disconnect between rational preference and rational choice does not seem especially coherent (Portmore 2007, 50). Yet rule consequentialism is surely a coherent (if mistaken) view. What has gone wrong?

Most naturally, when the rule consequentialist prohibits the 'best' or 'value-maximizing' act, they do not really mean that the prohibited act is desirable all things considered, but only *antecedently* desirable, i.e. before we consider the distinctive reasons for desire that derive from an act's deontic status as morally right or wrong. In this sense, their initial value theory is inconclusive or incomplete.[23] It accounts for only *some* of our reasons for desire: agent-neutral welfarist reasons, perhaps. But these reasons for desire are not decisive. Let's unpack how this might work.

Rule consequentialists first identify the rules that are best in terms of impartial welfare (or what's antecedently desirable), and then specify that we have decisive reasons to act in accordance with these rules. Finally, they might add, we have overriding reasons to desire that we so act. This way, a prohibited act may be 'best' according to the antecedent (agent-neutral welfarist) reasons for desire, and yet be bad (undesirable) all things considered. This avoids the incoherence mentioned above. But it also brings out how convoluted the view really is. It is recognizably consequentialist in the sense that it takes (*some*) reasons for desire as fundamental, and subsequently derives an account of reasons for action. But then it goes back and "fills in" further reasons for desire — trumping the original axiology — to make sure that they fit the account of right action. In this sense it exhibits a deontological

[23] Thanks to Michael Smith for suggesting this interpretation.

streak: reasons for action are at least *partly* prior to reasons for desire. In other words, the initial axiology includes only some values (the 'non-moral', agent-neutral welfarist ones), and what's *right* serves to determine the remaining ('post-moral', all things considered) good.

On this analysis, the distinction between deontological and consequentialist moral theories is a matter of the relative priority they assign to reasons for action and reasons for desire. This is an updated version of the traditional idea that deontologists take 'the right' as prior to (and partly determinative of) 'the good', whereas consequentialists think we can explain what's right entirely in terms of promoting non-moral goodness. Though the right and the good are clearly connected in some way, we may wonder which is more fundamental: which explains, or is the basis for, the other.

One appealing aspect of this analysis is that it allows us to reject recent arguments from the possibility of agent-relative value to the conclusion that really *all* theories are consequentialist (cf. Louise 2004). Critics of the distinction have noted, for example, that a deontological prohibition on lying may prove extensionally equivalent to a consequentialist injunction to maximize the agent- and time-relative value of *being yourself now honest*. But we are not thereby forced to collapse the distinction between deontology and consequentialism, so long as we can distinguish the two possible ways the relative priority might work out. The consequentialist version of the view would say that one has reason to act honestly *because* of the antecedent desirability of being oneself honest. A deontologist, by contrast, might claim the reverse: that the desirability of being oneself honest instead derives from the independent rightness or choiceworthiness of such acts. I think it is important to maintain this distinction. For while some agent-relative views, e.g. ethical egoism, may be naturally understood as goal-directed or consequentialist in nature, it seems a distortion to suggest that traditional deontological views are really just about promoting the goal of one's own

moral purity. (Even if that accurately describes their *upshot*, it does not seem not a fair characterization of what deontological theories *present* as morally significant.)

1.6 Does Fitting Choice Collapse into Desirability?

We are now in a position to address what I see as the most pressing objection to the argument of this chapter. I've been arguing that the fittingness view allows us to expand the expressive power of our normative theorizing, by raising new questions of 'fitting choice' over and above questions of value or desirability. But one might object: *why think that there are further normative facts about 'fitting choice', over and above the facts about fitting desire?*[24] To bring out the problem, let us imagine that an agent is faced with a choice between two actions: she may either ϕ or ψ. Further suppose that she has determined that ϕ-ing is the option that is all things considered most desirable. Is there any further question here of what she should choose? It may seem not. Again, it would seem somehow incoherent to think that ϕ-ing is the most desirable action, and yet that she ought to ψ instead. Choosing to ψ is clearly in tension with wanting, all things considered, to ϕ; it seems that one or other attitude must be mistaken (inappropriate, unfitting). This may be taken to suggest that questions of fitting choice are reducible to questions of fitting desire, so that there is not really any 'further fact' here of the sort that I have suggested.

I have three responses to this objection. Firstly, as discussed in footnote 14, there may be cases like 'Buridan's ass' in which an agent must choose between equally desirable options. In such a case, it seems that either choice (of one over the other) would be fitting, though it would not be fitting to prefer either option over the other. This suggests a subtle difference between the norms that govern preference and those that govern choice, albeit one that may only show itself in a very limited class of situations.

[24] Thanks to Michael Smith for pressing me on this.

Secondly, even in the remaining cases, I do not think that it is strictly *self-contradictory* to claim that what it is fitting to choose diverges from what it is desirable to do. Instead, like the claim *that suffering is good*, or *that only the welfare of a certain race of people matters*, I take it to merely be a very obvious but non-analytic falsehood. One could hold the contrary view without contradiction; it just isn't remotely plausible as a matter of substantive normative fact. So our imagined agent, having determined that it's most desirable that she ϕ, does in fact face a further question of whether she should choose to ϕ. It merely happens to be a very easy question.

Thirdly, even if one thought that fitting choice and fitting desire must coincide as a matter of conceptual necessity, no conceptual *reduction* is possible if the direction of explanation remains a conceptually open question, as I think it clearly does. Although act consequentialists think that acts are choiceworthy in virtue of being all things considered desirable, my previous section argued that we should understand rule consequentialists and deontologists as reversing the order of explanation. In the case of rule consequentialists, we find that some things (e.g. aggregate welfare) are indeed antecedently desirable, but then facts about what's choiceworthy intervene to influence what is ultimately all-things-considered desirable. Deontologists might even go so far as to take the facts about choiceworthiness as wholly fundamental and determinative of what's desirable. Now, I take it to be a matter of substantive normative fact which of these is the true moral theory. It would be prejudicial for our conceptual framework to take a stand on this issue. So we should acknowledge that there is room in our normative theorizing for both 'fitting choice' and 'fitting desire' assessments, and that it is a matter of substantive normative fact which of these is prior. There are two distinct conceptual possibilities here, as there would not be if the distinction between fitting choice and desirability of choice were to collapse. So the distinction must be upheld.

One might object that my earlier arguments against Global Consequentialism show that the Fittingness view itself violates the desideratum of neutrality between first-order normative theories.[25] But I think my analysis is better understood as showing that Global Consequentialism is not a purely first-order normative view (in the relevant sense). As we saw in section 1.4, we can decompose Global Consequentialism into two components: a first-order axiology that might be shared by any other form of consequentialism, and the structural claim ('structural' in the sense found in our third desideratum from §1.1.1) that *the normative domain is flat and symmetrical in structure, such that judgment-sensitive and non-judgment-sensitive evaluands are on a par*. It is only this latter claim that I object to.[26]

Recall the motivation for neutrality: we wish to select a conceptual framework within which to conduct our first-order inquiry. Even if it were a count against the Fittingness view that it rules out Global Consequentialism, that's clearly still a huge improvement over the value primitivist's exclusion of *every* form of non-consequentialism. But given my analysis of Global Consequentialism in this chapter, I don't think we should really see it as a cost at all. We've seen that, so far as its first-order normative claims are concerned, Global Consequentialism does not make any further claims beyond Act Consequentialism.[27] And the fittingness view is neutral with respect to *those* claims. It is only the Global Consequentialist's assertion of the parity thesis that is rejected by the Fittingness view. But that's not a problem, because one of the desiderata for our conceptual framework is to settle this 'struc-

[25] Thanks to an anonymous referee for suggesting this objection.

[26] And even this is not strictly incompatible with the Fittingness view alone. As explained in section 1.3, the Fittingness view provides a formula for determining the scope of normativity: there will be a kind of normative status corresponding to each kind of rational output. So the result that there are other normative assessments to be made besides assessments of desirability only follows from the Fittingness view once it is supplemented by the further (albeit obvious) claim that there are other rational outputs besides desires.

[27] If anything, it may — as we've seen — make fewer claims, if in calling an act 'right' they are *merely* saying that it maximizes value, and not further claiming that it thereby has the distinct property of being fitting to choose.

tural' question about the natural joints of the normative domain. We did not *want* to remain neutral on that.

Another way to make the point is that adopting value primitivism as one's conceptual framework effectively just begs the question against non-consequentialist views. My arguments for the Fittingness view, by contrast, have (if successful) *shown* why the structural elements of Global Consequentialism are mistaken. I have not simply assumed it. So I do not think the Fittingness view is objectionably prejudicial in this regard.

Conclusion

In this chapter, I have argued that we should take *fittingness*, rather than *value*, as our normative primitive. This allows us to say everything that we could have said using value-talk, and more besides. It helps us to avoid the misguided global consequentialist idea that judgment-sensitive evaluands (e.g. choice or action) are 'on a par with' — or normatively assessable in all the same ways as — mere evaluands such as eye colours. Instead, we find that there's a distinct property of rightness that applies only to acts, over and above the property of value (or desirability) that all sorts of evaluands might have. Finally, the fittingness view, unlike value primitivism, provides an appropriately ecumenical conceptual framework that remains neutral on the first-order normative dispute between consequentialism and deontology.

Chapter 2

The Fitting and the Fortunate

Critics of consequentialism often object to *how a consequentialist agent would (allegedly) think*. They claim that the consequentialist agent is, in some sense, a *bad* character: cold and calculating, alienated from themselves and others, etc. Defenders of consequentialism typically dismiss such objections by citing the distinction between 'criteria of rightness' and 'decision procedures'. (Utility provides the criterion that determines the moral status of an act, but it's a further question whether agents ought to attempt to calculate utilities themselves. If it'd have bad results then consequentialists would recommend against it!) However, I argue that the strongest objection in this vicinity is not just that thinking like a consequentialist would have bad results, but that such a psychology would be morally *perverse*, in a sense that's incompatible with the psychology in question constituting a morally *accurate* way of thinking. I then assess several instances of this argument form, with particular attention to the 'value receptacle' objection.

Introduction

In order to assess the objection that the consequentialist agent would be of bad character, we must distinguish two very different kinds of normative evaluation: the

'fortunate', and the 'fitting'.[1] If a psychological trait is *fortunate*, or instrumentally valuable, then it will be recommended by the normative theory as something to aim *at*. Roughly, this is to say that it is desirable, or ought to be inculcated. For example, if caring intrinsically about your friends, or your stamp collection, is conducive to your own happiness, then egoistic hedonism would evaluate such concerns positively in this sense. But there's another important sense in which these are not properly *hedonistic* concerns. Reflecting this second kind of normative evaluation, we can ask whether the agent's psychology *fits with* or *embodies* the perspective of the normative theory — roughly, whether the agent is responsive to the reasons posited by the theory, in the sense that he desires what the theory says is desirable (one's own happiness, in case of hedonism), and has otherwise fully internalized the truth of the theory. This is to ask whether the agent's psychology is, in a sense, appropriate, rational, virtuous, or (as I will say) *fitting*.[2] Just as we make absolute assessments of rationality (i.e., rationality "according to the true theory") in addition to assessments of what's rational according to any particular theory, so we can assess whether a psychology is virtuous or morally fitting *tout court*, or fits with the true moral theory, in addition to assessing whether it fits with any particular theory. Both "theory-relative" and "absolute" fittingness judgments will prove important in what follows.

What's fitting and what's fortunate may come apart, as in the 'paradox of hedonism': one is more likely to achieve hedonistically-fortunate results (i.e. happiness) if one does not possess the hedonistically-fitting mindset of adopting happiness as one's supreme goal. For another example: agent-neutral consequentialists might argue that things actually turn out better when people have more partial motivations (Jackson

[1] This mirrors the familiar distinction between state-given and object-given reasons for attitudes. See, e.g., Parfit (2011, 50).

[2] I assume that consequentialism aspires to claim the full normative authority of practical reason, or the all-things-considered 'ought'. But those who prefer to understand the view as *merely* concerning some rationally-optional 'morality' are free to read me as simply bracketing whatever other, non-moral considerations may intrude. My subsequent talk of 'rationality' should thus be understood as concerning just *what's rational from the moral point of view*.

1991), and so this is something we should want to encourage; but that doesn't change the fact that impartiality is really the more fitting or morally *accurate* perspective, according to their theory. It's just to say that some things are more important than being virtuous, or believing and fitting one's psychology to the truth.

With this distinction in hand, we can now identify two interpretations of the "bad character" objection to consequentialism. One claims that consequentialism is 'self-effacing' in the sense that it'd be *unfortunate* to possess the fitting consequentialist psychology. This is easily seen to be a poor objection. After all, it's always possible that events may conspire to punish those who are disposed to believe and act on the truth. Such circumstances do not cast doubt on a theory's claim to truth — again, it merely means that it may be better, from a moral perspective, if we believed some false moral theory instead.[3]

However, when deontologists complain about the bad character of a committed consequentialist agent, there is a stronger objection that they may wish to present. This is the objection that the fitting consequentialist psychology is (contrary to the consequentialist's claims) *not actually morally fitting*. For example, they argue that the consequentialist agent is incapable of friendship or commitment to projects — but, they may add, this seems like an intrinsic defect: surely *genuine* virtue and rationality are not incompatible with these important goods. So, they conclude, the consequentialist's conception of rationality (virtue, fittingness) must be in error. Note that it's no response to this to say that consequentialism doesn't necessarily recommend that one try to become rational or morally fitting in this way. For the objection we're now considering is *not* merely that it would be inadvisable for agents to be ideally rational or fitting. That isn't the problem. Rather, the objection is that consequentialism implies something *false* about what the ideally rational or morally fitting mindset would *be* — whether we should seek to adopt it or not.

[3] A more sophisticated version of the self-effacingness objection is explored in chapter 3.

This objection is the real challenge, though it risks being obscured if we focus exclusively on evaluations of fortunateness. It's important to note that our moral theory also commits us to a conception of the morally or rationally *fitting* agent. This creates space for critics to question whether our conception of the fitting agent is really accurate. (Given the connection between value and fitting desire, such objections will prove equivalent to challenging the consequentialist's value theory *insofar* as the objections are concerned with fitting desire, but as we will see in §2.4, some characterological objections are broader in scope, and so can't be reduced to axiological objections.) To take up this challenge, we must either (i) bite the bullet and insist that what the deontologist identifies as intrinsic moral 'defects' are not really so, or else (ii) argue that, properly understood, the fitting consequentialist agent would not in fact possess the identified defects.

In this chapter, I begin the latter task of rehabilitating our conception of the fitting consequentialist agent. In section 2.1, I identify and set aside a class of common objections that are only relevant to a peculiar subset of consequentialist views, so that we can turn our attention to characterological objections that purport to identify more general problems for consequentialism. Section 2.2 addresses various interpretations of the traditional 'value receptacle' objection. There I argue that it's a misconception to think that the fitting utilitarian agent would treat an individual's welfare as a fungible mere means to the end of aggregate welfare. Section 2.3 examines two influential objections from Bernard Williams, concerning whether a consequentialist agent would think either too much or too little about certain decisions. Finally, in section 2.4 I address the general worry that consequentialism embodies a perversely 'objective', mechanical attitude that would prevent the consequentialist agent from properly relating to other persons. In all of these cases, I argue, the fitting consequentialist perspective does not in fact exhibit the perverse feature being attributed to it.

The objections thus surveyed are familiar ones. But we will find that it is illuminating to recast them in terms of fitting psychologies. There are three main reasons for this: One is that it allows for greater precision in specifying the objection and identifying its strongest form (this comes out most clearly in the discussion of replaceability in §2.2.2). Secondly, when the objection is thus clarified, it may suggest a natural new response. Finally, we may (in some cases) find ourselves with stonger intuitions about the moral appropriateness or perversity of concrete psychologies. So when long-running debates over abstract moral principles reach a dialectical stalemate, recasting them in a new light may allow more progress to be made.

2.1 Axiological Refinements

Before we begin, let me first set aside a class of objections that *aren't* relevant to the assessment of consequentialism as such, namely, those which trade on particular assumptions about *what* is of value (rather than how we should respond to whatever is of value). For example, we might consider it perverse to plug into Nozick (1974)'s experience machine, but if so, the lesson to take from this is not that consequentialism is wrong, but merely that the goods to be maximized include non-experiential goods. Similar lessons may be drawn from 'Peeping Tom' objections: even if the victim never finds out about it, and hence experiences no pain or humiliation, we may still think that invasions of privacy *themselves* constitute a harm to the victim.

Axiological refinements also allow consequentialists to avoid the unpalatable 'mob rule' element of simple utilitarianism. It is entirely open to us to judge that sadistic pleasure, for example, has no positive value. So we need not think it appropriate for a sadistic majority to torture an individual, just because the quantity of pleasure thus obtained would factually outweigh the individual's pain. This larger quantity of pleasure may be of a nature that renders it *normatively* worthless.

Nor is this 'axiological' response to the problem merely an *ad hoc* move to save consequentialism. Rather, I think, anyone who considers this case a genuine problem to begin with should agree that the fundamental problem is one of disvalue rather than wrongdoing. We can see this by 'naturalizing' the scenario to replace the wrong action with a merely natural occurrence: Suppose that rather than a person torturing the minority individual, they were instead struck by lightning or some such. There is no action here, and so no wrongful action. Still, I have the strong intuition that this would be a bad thing to happen, no matter how much sadistic pleasure others might gain from their knowledge that this stigmatised individual suffered this harm. It would be very odd to think that sadistic torture was wrong, and yet that it would be a grand thing for the world if more stigmatised individuals suffered similar natural harms when accompanied by similarly greater net pleasure for sadistic observers. Clearly the problem here, if there's a problem at all, is that increased sadistic pleasure does *not* strike us as a good outcome. But if that's right, then consequentialism will not tell us to bring about this outcome.

Not every appropriate axiological refinement can be identified via the above 'naturalization' test. This is because there may be intrinsic disvalue to particular acts or choices themselves: sometimes it is precisely the vicious or inconsiderate exercise of agency that contributes to the disvalue of an outcome. For example, imagine a doctor who secretly molests his patients while they're unconscious. We cannot really 'naturalize' this scenario without losing its key feature: If the doctor merely 'touches' the patient as a result of an involuntary muscular spasm, for example, then this no longer qualifies as a violation in quite the same way, and hence it does not seem nearly as bad.

Once we allow that consequentialists may build the intrinsic disvalue of vicious action into their axiology, we need a more sophisticated test to distinguish them from deontologists — and hence to distinguish genuinely anti-consequentialist intuitions

from mere axiology-refining intuitions. At this point we may turn to agent- and time-neutrality. For while a consequentialist may be concerned to prevent vicious actions, she is ultimately no more concerned with her own actions than with other people's.[4] She will thus consider it worthwhile to perform a single intrinsically bad action herself if this prevents multiple similarly bad actions from others. And it is much more difficult for the opponent of consequentialism to establish the intuitive repugnance of *this* than to poke holes in crude axiological theories like value hedonism.

In section 2.2, we will explore some non-axiological worries about 'replaceability' or 'value receptacles' that are more plausibly objections to consequentialism *as such*.[5] But before we do, let us first set aside two versions of the 'replaceability' worry that are merely axiological in nature.

2.1.1 Death and Replacement

One might worry about the fact that classical utilitarians attribute no significance to the ways in which experiences are packaged together into distinct lives, and hence only see death as bad insofar as it causes there to be fewer good experiences in future. We may worry that this does not do justice to the badness of death: most of us would not think it a good thing (all else equal) for someone to be struck down in the prime of life and replaced with a marginally happier substitute. The premature death of an individual is bad in a way that goes beyond the mere failure to create future goods. Death is not equivalent, as this view would have it, to the failure to create life.

[4] Here I assume that we are talking about *agent-neutral* consequentialism. Some agent-relative theories, e.g. egoism, are also plausibly consequentialist in nature, but then we can repeat the test using temporal neutrality rather than agent neutrality. Any theory that builds in time-relative 'side constraints' that should not be violated now even to prevent more such violations in future, sounds fundamentally deontological to me. Even if one could model the theory using time-relative values to be maximized, as in Louise (2004), this seems to distort the motivation for the theory. See chapter 1 for more detail.

[5] We will see that there is a subtle sense in which even the later objections qualify as 'axiological', but they will concern the general *structure* of our value theory rather than its particular *contents*.

But we can accommodate this intuition without abandoning consequentialism. We merely need to refine our axiology so as to properly capture the disvalue of death. Here's one possibility: Besides preventing the creation of future goods, death is also positively disvaluable insofar as it involves the interruption and thwarting of important life plans, projects, and goals.[6] If such thwarting has sufficient disvalue, it could well outweigh the slight increase in hedonic value obtained in the replacement scenario. Consequentialists are thus fully able to attribute significance to the packaging of experiences into lives, and to acknowledges the positive disvalue of death — they just need the right theory of value.

2.1.2 Imprecise Values

One might also object to (an implication of) the traditional consequentialist practice of assigning exact numerical values to things.[7] Suppose we begin with two people, neither of whom has a more valuable life than the other, and you can save only one. It doesn't seem that mildly "sweetening" one of the options, with a dollar bill or the like, should break the tie or make the choice any easier or less arbitrary. Consequentialists may accommodate this phenomenon of *resistence to sweetening* by — once again — appropriately complicating their value theory. Rather than holding the two lives to be precisely equal in value, they must be merely *roughly* equal (Parfit 1984, 431), or 'on a par' (Chang 2002), such that sweetening one option does not necessarily make it of greater total value than the other (despite being better than it was prior to sweetening).

While there is some intuitive support for the thought that resistence to sweetening is often appropriate, I don't think that it would be at all *immoral* to insist on

[6] Importantly, this account of the positive badness of death avoids the opposite mistake of attributing *constant* and *unconditional* disvalue to death. There may be circumstances in which death is an unmitigated blessing, after all. Instead whether — and to what extent — death constitutes a positive harm will depend on the situation, i.e. what important life projects it cuts short.

[7] Thanks to Daniel Greco for bringing this to my attention.

precise values.[8] As we will see, it's a mistake to think that treating people's lives as *comparable* in value entails treating them as *fungible* or interchangeable in the way that we treat money, for example, as being. I might be genuinely torn between two distinct but equal intrinsic values, recognizing the separate force of each, even as my decision hangs in the balance such that the slightest inducement to either side would sway my decision. The sensitivity of my decision to further incentives does not in any way imply a failure to appreciate the distinct and irreplaceable conflicting values in play. So the separateness (or non-fungibility) of values cannot be understood merely as a matter of their being not precisely comparable. We need a better account. In the following sections — most notably §2.2.2 — I advance a positive account of what it really takes to disrespect a person by treating them as fungible, and how consequentialists can avoid this fate.

2.2 Value Receptacles

One traditional objection to consequentialism — expounded, for example, in (Regan 2004) — is that it constitutes a perspective from which individuals are seen as mere 'receptacles' or repositories for whatever happens to be of value: let's stick with happiness, for simplicity. The general worry is that a genuinely consequentialist agent would fail to recognize other individuals as valuable ends in themselves; instead, the objection goes, individuals are seen as merely *instrumental* to the end of realizing happiness (or value more broadly) in the world. This objection may be refined in a couple of different ways.[9] Let us consider them in turn.

[8] While perhaps a "moral error" in some abstract sense, it is not *disrespectful* of another's person in the sense discussed in later sections.

[9] Thanks to Pablo Stafforini for helpful discussion on this point.

2.2.1 Incidental Interests

It's widely agreed that we have reasons to help other people. But we may ask about the deeper structure of these reasons: *why* do we have this reason? On whose behalf does this reason exert its normative force or make claims on us? The commonsense answer is that these normative reasons speak on the behalf of the individuals who need our help. It is *for their sake* that we have reason to relieve their suffering. This much seems clear.

Yet utilitarians might be thought to deny this datum. As Singer (1993, 121) puts it, "The total version of utilitarianism regards sentient beings as valuable only in so far as they make possible the existence of intrinsically valuable experiences like pleasure." There is no mention here of the interests of the beings experiencing these pleasures. If the utilitarian's theory simply tells her to maximize net happiness, it may seem natural to reconstruct the fitting utilitarian's thought-process as follows: *Bob is in agony. My goal is to maximize utility, i.e., the balance of pleasure over pain. There is some agony (namely, Bob's) that I am in a position to relieve. Doing so would serve my goal. So I will act to relieve Bob's suffering.* But now note that the interests of *Bob himself* seem to have dropped out of the picture for our imagined utilitarian agent. She is merely concerned to minimize pain and suffering. The fact that doing so is *good for Bob* (or anyone else) is not a relevant consideration to her way of thinking, or so we might imagine. Helping people is incidental, a mere side-effect to her real goal of patterning the universe with a particular class of experiences. Call this view *Utility Fundamentalism*.

By taking the value of pleasure (and disvalue of pain) as fundamental, and not to be explained in terms of their value *for* individuals, Utility Fundamentalism seems objectionably fetishistic. It treats individuals as intrinsically valueless 'receptacles', of moral interest only insofar as they provide a space or habitat for what (supposedly)

really matters: the brute promotion of pleasure over pain. This moral perspective strikes us, I think rightly, as perverse.

If this is how we are to understand the 'value receptacle' objection, then utilitarians (and consequentialists more broadly) may escape it simply by rejecting Utility Fundamentalism. After all, there is a very natural alternative account, according to which pleasure (say) is good precisely *because* it is good *for* the individual who experiences it, and suffering is bad because it is bad for the suffering individual (Wilson 2006). On this view — call it *Welfarism* — the interests of individuals play an essential explanatory role in our value theory. When the welfarist utilitarian relieves Bob's suffering, the fact that this benefits Bob is not merely incidental to her reason for acting. It is, on the contrary, the source or ground of her reason. She has reason to relieve suffering precisely because this is good for someone.

We may demonstrate the difference between these two views by way of a fanciful counterfactual: If the welfarist utilitarian became convinced that some pain was, for some reason, intrinsically good for Bob,[10] she would no longer take herself to have non-instrumental reason to rid Bob of it. The utility fundamentalist, by contrast, has a fixed goal that makes no mention of the interests of individuals *as such*. She cares about experiences, not experienc*ers*. So even if she too believed pain to be good for Bob rather than bad for him, this would be of no intrinsic interest to her: she just wants to minimize pain, no matter whether this helps or harms the individuals experiencing the pain in question.

We thus see that only the utility fundamentalist is liable to the 'value receptacle' objection, understood as the failure to recognize that happiness (or whatever) is good just because it's good *for* individuals. This fetishistic perspective is by no means endemic to consequentialism. Indeed, it is entirely natural for consequentialists to instead take the welfarist route of specifying that happiness is good precisely

[10] Note that while this belief may be necessarily false, it's possible for someone to fail to realise this, and so to sincerely hold this misguided belief.

because it's good for the individual who experiences it. Our current interpretation of the value receptacle objection is then simply inapplicable to this welfarist form of consequentialism.

2.2.2 Are Persons Replaceable?

Even given an appropriately welfare-based explanation of why happiness matters, there remains a second interpretation of the 'value receptacle' objection that might be leveled against the utilitarian. The remaining objection is that utilitarians treat particular individuals not as ends in themselves, but merely as fungible or replaceable means to the end of promoting *aggregate* welfare.

This objection has been formulated in several different ways. Rawls (1999, 24) famously objected that "Utilitarianism does not take seriously the distinction between persons." Singer (1993, 121) writes, "It is as if sentient beings are receptacles of something valuable and it does not matter if a receptacle gets broken, so long as there is another receptacle to which the contents can be transferred without any getting spilt." The common thought here is that there's an important sense in which utilitarianism fails to treat us *as* individuals. It takes our interests into account, perhaps even *as* interests, but not in a way that appreciates the normative *distinctness* of my interests and yours. We are all melded together, into a kind of unstructured, undifferentiated welfare soup.

These formulations are evocative, but imprecise. I think we get a firmer grip on the objection by formulating it in explicitly psychological terms. The fitting utilitarian agent would (allegedly) have but a single ultimate desire: to maximize aggregate welfare. They thus see different individuals as interchangeable. It makes no difference, to such an agent, which of several people is helped (or indeed whether one person is helped a lot or several people each helped a little), so long as the impact on aggregate welfare would be the same in either case.

To bring out why this is so objectionable, note that fungibility is, in general, the mark of the instrumental. Money is fungible precisely because we do not value the possession of *particular* bills: replacing two tens with a twenty would serve my ends just as well. For another example, if my sole ultimate desire is to slake my thirst, then I will be indifferent between two equally effective means to satisfying this goal. If someone switches my glass of water for another that's qualitatively identical, this is not a change that's normatively significant to me. I do not desire *that* glass in particular, so it may just as well be replaced by any other that would do the job. On the other hand, if I *had* (bizarrely) desired the original glass in its particularity, then the substitution would be of significance to me: it would thwart one of my non-instrumental desires.

This connection between fungibility and merely instrumental valuation explains why the above objection to utilitarianism seems so forceful. It seems perverse to treat individuals as replaceable or fungible, because such treatment constitutes a failure to intrinsically value individuals in their particularity. The correct moral theory, we feel, must attribute intrinsic value to particular individuals and not just to the general welfare.

How is a theory to satisfy this requirement? Again we can clarify the matter by reference to fitting psychologies. We have seen that it's morally perverse for an agent to be *indifferent* between options that equally benefit distinct people, for that is to disrespect the individuals by treating them as fungible means to the aggregate welfare. But of course we do not want to favour either person over the other, since such bias would constitute disrespect for the person whose equal benefit we counted for less. Instead, I propose, the fitting response to a tradeoff between two distinct but equally weighty values is to feel *ambivalent* about the choice. There are distinct reasons pulling you in either direction, corresponding to the distinct values served by

either choice. But these reasons are equally weighty, so the agent is *torn* rather than pulled without resistance towards one choice over the other.

This is a distinction we should want our theories to be able to make. Whatever substantive disputes we may have about what is of value, we should all acknowledge the formal difference between (i) a pair of options serving distinct but equally weighty final values, and (ii) a pair of options that serve literally one and the same value. For example, assuming that token artworks have intrinsic value, a choice between saving a great painting or an equally great sculpture is importantly different from a choice between saving the same painting in either of two different (but equally effective) ways. In the latter case, the two options are seen to serve the same token value in virtue of saving the same token artwork. Other cases of this may be more subtle, as even two distinct concrete objects may serve as vessels for one and the same token value. An intuitive example of this is pleasure: I'm completely indifferent between the prospects of a massage for my left foot or my right, assuming that either would be similarly pleasant.[11] I take this to suggest that left-foot-pleasure and right-foot-pleasure are not distinct final values, the way that the painting and the sculpture (or my welfare and your welfare) are. Instead, it seems, I ultimately value pleasures of a certain qualitative kind *in the aggregate*, and particular instances of such pleasures are thus, in an important sense, of merely 'instrumental' value to me. Of course this is not to say that they are causally instrumental to some downstream effect. We may instead call it *constitutive instrumentality*, as each token of pleasure is a constitutive, rather than causal, means to the end of aggregate pleasure.

With this understanding in hand, we may now characterize the replaceability objection as alleging that fitting consequentialists must likewise treat *individual persons* as constitutive means to the aggregate welfare, rather than as distinct ends in themselves. Given that individual persons have final value, such instrumental treatment

[11] Thanks to an anonymous referee for prompting me to discuss this case.

constitutes a distinctive kind of *disrespect* or failure to respond appropriately to the value that persons have in themselves.

The reader should now have an intuitive grasp of the distinction between (equally-weighty) distinct final values and (equally effective) mere means to a single final value. I've suggested that one way this distinction might play out is that in the second case the two options are perfect substitutes, and hence the fitting attitude for an agent to take towards them is indifference. In the former case, by contrast, the two options are not *substitutes*; they serve different ends, albeit equally worthy ones. This naturally suggests that the fitting attitude to take is ambivalence, rather than indifference.

Another way to support this conclusion is via the idea that it's fitting to intrinsically desire *each* intrinsic good, with strength proportional to the magnitude of the object's value. If, and only if, a pair of options serve distinct intrinsic values, will the two options differentially satisfy the intrinsic desires of the morally fitting agent (and hence strike her as significantly distinct). Insofar as the agent has conflicting desires, we can say that she manifests ambivalence rather than indifference over the options.

We are now in a position to evaluate the objection that utilitarianism treats people, and their interests, as fungible. This is, as we have seen, equivalent to interpreting utilitarianism as the view that only one token thing, namely aggregate welfare, has intrinsic value. Call this view *monistic utilitarianism*. This view really does neglect the separateness of persons, for it attributes intrinsic value merely to the whole, and not to each of us in our particularity. As a consequence, the fitting monistic utilitarian has but a single desire — to maximize welfare — and treats our individual interests and concerns as mere (constitutive) means to the satisfaction of this more global goal. This is, I agree, morally perverse.

But there is no reason why utilitarianism must take this monistic form. There is a very natural alternative view, call it *pluralistic utilitarianism*, on which *each*

particular person's interests are (separately) accorded final value.[12] There is not just one thing, the global happiness, that is good. Instead, there is my happiness, your happiness, Bob's, and Sally's, which are all equally weighty but nonetheless distinct intrinsic goods. What this means is that the morally fitting agent should have a corresponding plurality of non-instrumental desires: for my welfare, yours, Bob's, and Sally's. Tradeoffs between us may be made, but they are acknowledged as genuine tradeoffs: though a benefit to one may outweigh a smaller harm to another, this does not *cancel* it. The harm remains regrettable, for that person's sake, even if we ultimately have most reason to accept it for the sake of more greatly benefitting another.

Contrast this with the case of money: If you have to invest $5 to earn $10, there is nothing to regret. The $5 is a "cost" merely in the sense that it would have been *even better* if you could have attained the $10 payoff without having to pay the $5. But given that this is not an option, there is nothing regrettable about the deal *as a whole*, the way that there is something regrettable about benefitting one person greatly at lesser cost to another. We can explain the difference, in the cash case, as a matter of both sums of money being mere components or constituents of the single token value, or desirable end, of aggregate wealth. This is very different from how the pluralistic utilitarian conceives of welfare tradeoffs between distinct persons.

We thus find that (pluralistic) utilitarianism is well able to reflect the normative separateness of persons, and to avoid treating people as fungible, replaceable receptacles of value. This is, if correct, an important result: It's commonly thought that the utilitarian's willingness to weigh harms to one person against benefits to another essentially involves treating the one as a "mere means". But my above analysis suggests that this traditional thought is simply confused. One may have thoroughly

[12] The view may still be monistic in the sense that there's just one *type* of thing that's good (cf. Hurka 1996). But the crucial point for present purposes is that there are a plurality of *token* final values. The separateness of persons merely requires that we each be valued separately. There's nothing obviously objectionable about it turning out that we are valuable in the same kind of way.

non-instrumental desires for each of two distinct instrinsic goods, and make reluctant tradeoffs between them in a way that is importantly different in kind from the tradeoffs one makes with fungible goods like money. The mere willingness to balance conflicting values is not itself constitutive of instrumental or fungible treatment. Critics may still insist that utilitarianism is just *extensionally incorrect* in its prescriptions for morally right action, but those wanting to make stronger claims about 'value receptacles' need to back up their claims with a rival account of instrumental valuation — as such rhetoric is seen to be baseless if the present account of instrumental valuation is correct.

An interesting implication of my account is that we may find that we actually treat our interests-at-a-time as fungible.[13] While we might initially have assumed that our momentary interests have final value, we may find on reflection that we consider our interests across time, unlike interests across people, to be properly fungible. As in the case of fungible pleasures, this view can easily be incorporated into my framework by positing that individuals' interests-at-times are mere constitutive means to the final good of their timeless welfare. Alternatively, you might opt for the view that it's fitting to consider tradeoffs between timeslices to be just as emotionally fraught as tradeoffs between persons, and so assign final value to each momentary self individually. For purposes of this chapter, I can remain neutral on this question of whether to attribute final value to momentary welfare, or only to timeless welfare.

2.2.3 Objections

I have argued that the fitting utilitarian could respect the distinctions between persons by separately desiring the good of *each* person's welfare, rather than having a single, totalizing desire for the aggregate good. But a difficulty arises when we consider goods that the agent is unaware of. Consider some particular unknown person, Harry. Our

[13] Thanks to an anonymous referee for bringing this to my attention.

utilitarian cannot have a particularized desire for Harry's welfare, since she cannot even refer to Harry in particular. But her values must extend to others somehow: It's not as though she'd accept an offer to improve the welfare of her neighbour Bob at greater cost to some unknown other. So it seems that we need something like a generic desire for aggregate welfare to step in and fill the gap. (To avoid double-counting, we'd probably need to exclude Bob — and any others for whom the agent already has a particularized concern — from the remaining aggregate.)

Is this a problem? Perhaps not. It doesn't seem so objectionable to treat people you've never even heard of as faceless members of the aggregate. How could they be *other* than faceless and generic to you? Moreover, the agent's attitude here is not merely instrumental. It's not as though our utilitarian thinks that unknown people fundamentally matter only in respect of their being members of the unknown aggregate. Rather, her concern for unknown people's aggregate welfare is a stop-gap measure that reflects, in the only way possible, her appreciation of the fact that each of those individual unknown persons fundamentally matters in their own right. She knows that, if she knew more, she would form particularized desires for the welfare of each; but in the absence of the requisite identifying information, the best she can do to respect these unknown values is to fall back on the generic desire for aggregate welfare, as a kind of placeholder.

So far, so good. But what about *merely possible* future persons? (Compare Parfit (1984)'s 'Non-Identity Problem'.) Here the placeholder strategy seems dubious. Before, we were holding the place for the particularized desires we would have if fully informed — and it seems reasonable for an agent to deferentially ascribe normative authority to her fully-informed desires. But in case of merely possible persons, the barrier to particularized reference is metaphysical, not merely epistemic: There *is* no such particular person to refer to. The most we can appeal to is the *counterfactual* desire that we (ideally) would have had if someone else had existed. We would

have formed a particular desire for that someone's welfare. But so what? As things stand, there is no such person, and hence no valuable entity for us to respect as best we can. We cannot have a 'second-personal' reason (Darwall 2006), grounded in the normative authority of the non-existent individual himself, to take his possible welfare into account. Our concern must instead be, in a sense, 'impersonal'.

Even so, this needn't return us to any single, totalizing desire that the world be thus-and-so. We may instead have distinct desires for each possible generic good. We still need to distinguish between indifferent and ambivalent pairs of prospects, after all. For example, I may desire both that Anne have a happy child rather than none at all, and that Beth have a happy child rather than none at all. Perhaps I cannot coherently desire these things *for* the sakes of the respective children (especially if they never actually exist), but I can desire them — for the sake of the world, perhaps. And in so doing, I recognize that the prospective persons are not fungible, in the following sense: Despite being of equal value, there is a morally relevant difference between a world where only Anne has a child, and a world where only Beth has a child. The comparison calls for ambivalence, rather than indifference, since they serve distinct (though equally weighty) ends or ideal desires. If Anne's child would have a better life, then I could prefer that she be the one to come into existence, even while I regret the absence of Beth's possible child, whose life would have been (distinctly) intrinsically valuable in its own right.

The objector might respond by suggesting that it's only because of the differential impact on the existing individuals Anne and Beth that we see a significant difference here. If we imagine some more thoroughly generic question — say, whether the 100th child born in the year 2500 is a boy or a girl — indifference may seem the only appropriate response. I'm not sure about this, as our lack of response may just be due to our contingent failure to really vividly appreciate what a significant intrinsic difference the identity of each individual makes. But even if the critic is right here,

it's not clear that this is any objection to consequentialism in particular. If it turns out that distant future people cannot *but* be thought of as fungible, in the noted sense, then this limitation will presumably apply to all moral theories.

So I think the objection ultimately fails. Even in the toughest case — that of merely possible persons, who cannot be the ultimate ground of our concern for their welfare — consequentialists can plausibly still desire each good separately, and hence refrain from treating people as fungible. And even if it turns out that I'm mistaken about this, and in fact merely possible persons *are* fungible, then that is no fault of consequentialism. It would instead be a constraint that any moral theory must work within. So the remaining challenge for a theory would just be to ensure that it doesn't inappropriately extend the domain of the fungible to include actually existing persons. Consequentialism can, as I have shown, meet this challenge.

2.3 Thoughts Too Many, or Too Few

Bernard Williams has objected that consequentialists would find some choices more obvious than they should be, and that there are other actions that they would inappropriately pause over. As an example of the first, consider *Jim and the Indians* (Williams 1973): Captain Pedro will kill twenty innocent locals, unless Jim elects to kill one of them. What should Jim do? Assuming that all else is equal, and Jim knows it, then agent-neutral consequentialism implies that he should kill the one. The mere fact that it is *he*, rather than Pedro, who does the killing here is irrelevant so far as choosing between the two options is concerned. I argue elsewhere[14] that rational heuristics for non-ideal agents might complicate the matter — as may our normative uncertainty, insofar as the truth of consequentialism is itself unobvious (Lenman 2004) — but in this chapter I'm focusing on the ideal case. And in the ideal case,

[14] See 'What's Fit for the Fallible', the third chapter of my dissertation.

where Jim has a full grasp of the situation and no cognitive limitations, I think the consequentialist should simply insist that the choice *would* (or should) be obvious.

Williams (1982) puts forward the trickier charge that consequentialist agents would have "one thought too many". When such an agent jumps into the water to save his drowning wife, Williams writes, "it might have been hoped... that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife." (p.18)

The objection here is presumably not that one *blindly* ought to save one's wife, no matter the consequences for others. That would not be plausible. Rather, I take it, the challenge is to clarify the proper *role* that moral considerations play in one's cognitive economy. In particular, we may need to distinguish between *reasons* and background *conditions* on the applicability of a reason (compare, e.g., Schroeder 2007, chapter 2). In the case at hand, we may think that the agent should be motivated *just* by a concern for his wife, though his acting on this motivation is contingent upon the fact that he isn't causing greater harm by doing so. I don't see any barrier to consequentialism taking this distinction on board. So just because the consequentialist agent *wouldn't* have saved his wife if it hadn't been permissible for him to do so, it doesn't follow that the fact of its permissibility entered into the content of his 'motivating thought'. The condition may have a merely *virtual* presence in the agent, in Pettit (1994)'s sense that the violation of the condition would *trigger* the agent's attention and reconsideration; but in the absence of this trigger it doesn't figure in his thoughts at all.

A second way to interpret the objection here is that it would seem a kind of moral fetishism for an agent to be motivated by considerations of moral duty *as such*, i.e. under that abstract description. But, as I argued in section 2.2.2, there is no good reason to interpret the fitting consequentialist agent as having such abstract motiva-

tions. Rather, we should understand them as having intrinsic desires corresponding to each concrete intrinsic good — such as the welfare of any given person. So understood, the consequentialist is moved directly by his concern for his wife, unmediated by any more abstract description of 'duty' or 'the general good', even while he remains sensitive to the interests of concrete others in a way that ensures he is not blind to duty or the general good.

A third worry suggested by this case is that the consequentialist is responding to his wife no differently from how he would respond to anyone else. A few things can be said about this. One is that, lacking omniscience, we are aware of the virtues of our loved ones in a way that we are not aware of strangers. So even an agent with a standing disposition to value all persons equally might well find the value and needs of people he is more familiar with to be more salient, and hence more emotionally motivating. More importantly, when we consider the ideal case, impartial consequentialists can offer an attractive response: though it may be psychologically impossible for us humans, ideally an agent really ought to care about *all* persons as intensely as we care about our loved ones. Finally, I should note that if one remains unattracted to impartiality as a moral ideal, there is always the option of adopting an agent-relative axiology, weighting the welfare of one's loved ones more heavily than that of mere strangers.

2.4 The Participant Attitude

Strawson (1962) famously distinguished two broad stances or attitudes we might take towards another person. One is the attitude of "involvement or participation in a human relationship", which might naturally give rise to such personal reactive attitudes as gratitude, resentment, and so on. The other, which Strawson calls the 'objective' attitude, he describes in section 4 of his paper as follows:

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; [...] to be managed or handled or cured or trained; perhaps simply to be avoided [...] If your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him.

We might summarize the distinction by saying that the objective attitude involves seeing the other as an object, or a force of nature, to be understood and perhaps manipulated towards desired ends, whereas the participant attitude involves seeing the other as a fellow agent — a co-participant, perhaps, in the collective project of living well.

This distinction may help us to elucidate the unease that many feel about consequentialism. For they may have a sense that consequentialism represents a distressingly mechanical or 'objective' view of the world, where one's fellow agents — and even one's own future self — are seen primarily as causal levers rather than as rational beings. Utilitarian discussions of punishment, for example, are centrally concerned with the mechanical question of how to bring about better outcomes. Consequentialists will generally have a lot to say about when it is or isn't useful to blame people, but tend to put aside questions of blame-*worthiness* as empty or misguided.

This exclusive focus on outcomes might seem to neglect some of the most important questions, for in our everyday relationships we are often more interested in what emotional responses would be *fitting* than in what responses would be *best*. It seems doubtful whether an agent who failed to share these concerns would even be capable of the sorts of genuine relationships we ordinarily (and, I think, rightly) prize. There's plausibly something *defective* (and not merely extrinsically unfortunate) about an agent who is constitutionally incapable of relating to others in this

way. So it is of the utmost importance for moral theorists to uncover whether the fitting consequentialist agent really *would* forsake the participant attitude in this way. For if so, this would provide grounds for a decisive objection to consequentialism: its conception of the 'fitting' moral agent would not be genuinely morally fitting at all.

I will address this challenge in two parts, first examining how consequentialist agents would relate to themselves, and second, how they might relate to others.

2.4.1 How we relate to ourselves

Some may worry that the consequentialist agent takes an objectionably 'objective' or mechanistic attitude, not just towards others, but even towards their future selves. Suppose that I am currently faced with a choice between A and B, and that tommorrow I will be offered a choice between two other options, C and D. Let us suppose that the value ordering of the possible outcomes is as follows: A & C > B & D > B & C > A & D. That is, my choosing A today could lead to either the best or the worst outcome, depending on whether I choose C or D tomorrow. Further suppose that I know, from past experience of similar choices, that I'm much more likely to choose D than C tomorrow, whatever I choose today. How should this fact affect my present decision-making?

According to 'Possibilists', it shouldn't affect my decision at all. It's within my power to choose A now, and C later, and so that's precisely what I should do. 'Actualists', by contrast, may admit that while A & C is the best available act-*sequence*, the decision facing me at the current moment is just what to choose *now*, and in light of the terrible expected value of now choosing A (since it will most likely be followed by my later choosing D), I should instead pick the safer option B (Jackson and Pargetter 1986).

Actualism is, I believe, the view that is more true to the spirit of consequentialism: We should act in the way that would maximize expected value, all things considered

— and one such thing to consider is our disposition to act wrongly in future. But one might object that this is to take an inappropriately alienated perspective towards one's future self. It is to treat one's future decisions as beyond one's (current) control, which may seem wrong-headed. This is how one might find it objectionable how the consequentialist agent relates to him or herself.

I think this objection is misguided. It is a factual question whether there is anything that the agent can do now (including mental activities like steeling her will) to ensure that she acts rightly in future. If there is something (A*) that the agent can now do to ensure this ideal outcome, then the Actualist will obviously join the Possibilist in recommending this option. On the other hand, if it's really *true* that whatever intentions she forms now won't make any (or enough) difference to her future actions, then it cannot be inappropriate for the agent to act in light of this truth. Indeed, it would seem plainly unwise for her *not* to! So I think our intuitive resistance to Actualism actually stems from imaginative resistance to the set-up: We tend to assume that, in choosing A, the agent *could* (concurrently) do something to secure her future compliance with the ideal plan. But we can only distinguish Actualist and Possibilist recommendations if we imagine a case where this ordinary assumption fails to hold. In such abnormal cases, I think the Actualist's prescribed action is the sensible choice for the agent. It really is appropriate to consider your future self as beyond your current control in those odd cases where this is true. On the other hand, Actualist Consequentialism does not recommend that one *always* adopt this alienated self-perspective. On the contrary, in all those ordinary cases where you can currently form intentions that will secure future right actions, that's precisely what you should now do![15]

[15] Doug Portmore has further developed this basic idea, which he calls 'Securitism', in chapter six of his (2011).

2.4.2 How we relate to others

We've seen that consequentialist agents would intimately identify with their future selves, when appropriate, by deliberating now about future actions. But this *deliberative* identification is obviously inapplicable to *inter*-personal relations. (We may decide how to act in future, but we cannot directly decide how others will act. At most we can decide what instructions, requests, or threats we will send their way.) So, in order to assess the force of the objection that consequentialists would not appropriately relate to others, we must first get clearer on what kind of interpersonal relation is appropriate.

In the interpersonal case, the relevant difference does not show up so clearly in what decisions we would make, but rather in what we would care about, and how we would respond emotionally to others' apparent attitudes towards us. As previously described, the participant attitude involves respecting the other person *as* a person, in a way that involves caring about whether they likewise respect us. When we adopt the participant attitude towards someone, we may respond with gratitude when they show a surfeit of good will towards us, and with resentment when they are inconsiderate or malicious. If we adopt the objective attitude, by contrast, we see the other from a much more detached and impersonal perspective, and are no more emotionally vulnerable to them than we would be to a valuable vase or a gushing waterfall. We may still want to protect these valued objects — the objection here is not that the consequentialist fails to value people. Rather, the objection is that one who exclusively adopts the objective attitude is failing to respond to other people in the appropriate *way*. Though we may in some sense be valuable objects, that is not *all* that we are. We are also rational *subjects*, a fact which calls for recognition from our fellow agents.

I'm not convinced that there is really any incompatibility between consequentialism and the participant attitude. Consequentialism as a theory is focused on a differ-

ent question — namely, how we should act.[16] But then the present objection merely suggests that consequentialism is incomplete, and we should seek to supplement this theory of fitting action with a theory of fitting reactive attitudes (gratitude, resentment, etc.). That is, a fitting consequentialist *may* fail to take up the participant attitude, but she need not so fail, and indeed if *fully* normatively fitting she would take it up when appropriate. Note that this is no more an objection to consequentialism within its domain than is the need to supplement it with an epistemological theory of fitting belief. There's more to normativity than just actions, and the right norms for other attitudes may differ from the consequentialist norms that are fitting for action. Consequentialists should, I believe, accept this fact about the limited scope of their theory.

One might press the objection by pointing out the practical incompatibility of the participant and objective attitudes. Insofar as the objective attitude — attending to the causal structure of the world, including its human inhabitants — better prepares one for consequentialist decision-making, and adopting this attitude is in practice incompatible with simultaneously adopting the participant attitude, then doesn't consequentialist decision-making effectively *preclude* one's sharing in the participant attitude? Here I find it illuminating to consider Bennett (ms)'s tentative suggestion for how to understand the practical tension between the participant and objective attitudes:

> Reactive attitudes essentially prepare for personal interaction of a certain kind, while the objective attitude prepares for inquiry, and these two sorts of activity are somehow incompatible. If that is right, the two sorts of attitude are derivatively in conflict, like simultaneously readying oneself for a sexual encounter and for giving an after-dinner speech.

[16] For a defence of this conception of consequentialism, contra Pettit and Smith (2000), see my first chapter, 'Fitting Attitudes for Consequentialists'. Let me also clarify that the consequentialist's axiology will have implications for what it's fitting to desire, but this still leaves open what the correct norms are for various emotional states, etc.

This account of the role of the two kinds of attitude seems plausible: There's something very relationship-oriented about the participant attitude, whereas the objective attitude seems forced on us when our task is to make an accurate prediction or accounting of things. But if we accept this, it allows us to dissolve the present objection, as there's no reason to think that agents (including consequentialist agents) must be incessantly inquiring. Admittedly, we often consider tricky cases of the sort where inquiry into the likely consequences of various possible decisions is indeed called for. So it is understandable that some would come to associate consequentialism with the objective attitude. But ordinary life does not call for constant inquiry,[17] so there is plenty of space left in the agent's life for the participant attitude to play its role.

Conclusion

In this chapter, I have distinguished two general forms that a character-based objection to consequentialism might take, mirroring the distinct normative assessments of what's *fortunate* and what's *fitting*. I suggested that objections from fittingness are more powerful, and that these objections are not adequately addressed by the standard consequentialist strategy of distinguishing criteria of rightness from decision procedures. I then took some initial steps towards formulating an adequate response to the fittingness objections against consequentialism. In particular, I showed that the consequentialist perspective does not involve seeing individuals as fungible means to the general good, that the consequentialist agent need not have 'one thought too many', and that they are not barred from adopting the 'participant attitude' that is a necessary prerequisite for interpersonal relationships.

[17] This is related to the problem of 'defective deliberateness' that I further discuss in chapter 3.

Chapter 3

What's Fit for the Fallible

What would a utilitarian agent look like? Some have taken the answer to describe an agent so incompetent and perverse that it casts doubt on utilitarianism itself. In this chapter, I develop the strongest form of this "self-effacingness" objection to utilitarianism, based on the idea of a constitutive link between rationality and *normally* competent agency. Assuming this understanding of rationality for sake of argument, I then suggest two ways to defend utilitarianism. One appeals to a Railtonian 'sophisticated' or indirect utilitarian psychology, though I suggest some potential problems for this approach. The second involves showing how we can develop a direct utilitarian psychology within rational constraints. In the course of distinguishing these two alternative paths, I make a distinction between dispositions that are 'extrinsically desirable' and those that are desirable in virtue of being 'well calibrated for action' — a distinction that I then employ to illuminate the Gauthier-Parfit debate about whether it's rational to act on rationally desirable dispositions.

Introduction

Bernard Williams (1973) famously objected that "utilitarianism's fate is to usher itself from the scene." The idea is that utilitarianism will tell us that we should try

to rid ourselves of our belief in it (and perhaps internalize some alternative theory instead), since this would better achieve the goals of the theory.

Taking Williams' worry as a starting point, this chapter will explore the relation between moral theories and morally fitting psychologies, with a particular focus on what's fitting for fallible, non-ideal agents. My main focus will be on utilitarianism, and whether the fitting utilitarian agent would be unacceptably incompetent in a way that casts doubt on the truth of the theory. But much of what I say in utilitarianism's defence could also be extended to Kantianism and other targets of Williams-style anti-theory objections.

I begin, in §3.1.1, by explicating the self-effacingness objection in greater detail, showing how it can be understood as challenging the utilitarian conception of a morally fitting agent, and why this 'fittingness objection' is a challenge to utilitarianism itself. §3.1.2 explains why the standard consequentialist response to character-based objections — namely, distinguishing criteria of rightness from decision procedures — is inadequate to meet the objection.

§3.2 explores the use of Railton (1984)'s 'sophisticated' consequentialist psychology in response to character-based objections. Insofar as the sophisticated consequentialist agent possesses whatever motivations would be most desirable, it may seem that what proponents of this account are describing is merely a *fortunate*, not necessarily any kind of *fitting*, psychology. We can imagine circumstances in which it would be most fortunate to possess positively malicious motivations, but that clearly wouldn't make malice a virtue (or 'fitting', as I use the term here). So, I will argue, this popular approach also fails to adequately address the fittingness objection.

In diagnosing where the Railtonian conception goes wrong, I suggest that we can identify a proper *subset* of desirable dispositions as 'well-calibrated' or *rationality-enhancing*, in contrast to others that merely cause better results by some external means. It is only the rationality-enhancing dispositions that we should attribute to

the fitting moral agent: Extrinsically desirable dispositions are worth pursuing, but once achieved may disqualify one's resulting psychology from being morally fitting. (Just as, for example, a hedonist should try to acquire non-instrumental interests of other kinds, but once they do they no longer qualify as fitting the theory of hedonism. They have rationally made themselves irrational, by hedonistic lights.)

As a brief aside, I take this distinction to illuminate the Gauthier-Parfit debate about rational transmission: i.e., whether it's always rational to act on a disposition that it's rational to acquire. In section §3.3, I argue that Parfit is correct that no such general principle holds: Even if the deterrence value renders it rational to acquire a doomsday disposition, it is not necessarily rational to *act* on such a disposition. Nonetheless, we may hope to vindicate a related version of the transmission principle that is restricted to those dispositions that are independently identifiable as rationality-enhancing or 'well-calibrated'.

Finally, in §3.4, I draw on this conception of 'well-calibrated' dispositions to show how I think the utilitarian can successfully respond to the self-effacingness objection. In doing so, I will draw heavily on assumptions presupposed by the objection: namely, that there are norms fit for human-sized (finite, fallible) minds, and that a person who meets these norms is thereby competent to act in a wide range of circumstances. My general strategy is thus to first specify some initial preconditions for competent human-like agency, and then to explicate how a recognizably utilitarian mindset might fit within those constraints. The resulting picture, if still not entirely attractive, may at least seem significantly less "defective" — and less likely to be typically self-effacing — than the standard caricature of a utilitarian agent.

To this end, I will first show that the utilitarian agent should not be conceived of as constantly engaging in deliberate calculation. This misconception may arise from assuming that every deliberative question to which there is an appropriate answer is thereby a question that it's appropriate *to ask*, but we will see that this cannot be so

for finite agents. Attention is a limited resource, and executive control should only intervene when by doing so it is likely to improve the quality of the agent's actions. *Excessive* executive control is, I will argue, no part of the fitting utilitarian mindset. So, the fitting utilitarian agent will not engage in predictably unreliable attempts at explicitly calculating utilities, but will rely more heavily (as many utilitarian philosophers have suggested) on the general rules of thumb that are more likely to see him right. In particular, I will argue that the fitting utilitarian will *not* be disposed to break a generally beneficial rule merely because the benefits seem to him to outweigh the costs. Even on straightforward act utilitarian grounds, his behaviour may be largely rule-governed in a way that renders him trustworthy and eligible for social co-operation. Here I do not *merely* argue, as others have done, that this is the most fortunate or utilitarian-recommended mindset (indeed, depending on empirical contingencies, it may not be). Rather, I will argue that such a mindset is *fitting* to the utilitarian's theory, in light of rational norms for fallible agents.

3.1 Self-Effacingness

3.1.1 The Objection

The self-effacingness objection, read straightforwardly, objects to the mere fact that utilitarianism is *in fact* self-effacing. But this cannot reasonably be thought objectionable. To see why, first observe that any plausible moral theory is at least *possibly* self-effacing in this way, because any plausible moral theory will tell us to acquire false moral beliefs if this is the only way to avoid absolute disaster — and it is always possible to contrive situations in which this is the very choice that we face. So it cannot be objectionable that utilitarianism is merely possibly self-effacing. Moreover, since the true moral theory is presumably non-contingent, it would be very odd to think that this objection suddenly gains additional force if we happen to live in one of

those possible worlds where the theory is self-effacing. Moral facts cannot plausibly depend in this way on our location in modal space. So the mere fact that a moral theory turns out to be *actually* self-effacing is not objectionable either.

Next we may note that utilitarianism is not *necessarily* self-effacing: there are possible worlds where it is (expectably) best to believe it. So one cannot form a sound objection from that premise. But perhaps it would be genuinely objectionable if utilitarianism were, in some sense, *normally* self-effacing.[1] And if, as critics charge, the utilitarian agent is constantly calculating, untrustworthy, apt to break generally beneficial rules whenever it strikes him as optimal to do so, etc., then we may indeed expect the view to be quite typically detrimental when internalized by fallible, human-like agents. The utilitarian mindset begins to look not just *unfortunate* ("given the circumstances"), but intrinsically defective.

This, then, is what I take to be the strongest interpretation of the self-effacingness objection. First, the critic assumes a strong connection between rationality and (metaphysically, not statistically) *normal*[2] competence, such that any normally incompetent agent is ipso facto lacking adequate rational capacities. Second, we need a link between moral agency and rational agency, such that virtue, or moral fittingness, is not incompatible with the possession of adequate rational capacities. Third, the critic posits that any fitting utilitarian agent would be normally incompetent. From this he draws his first conclusion: that the fitting utilitarian agent is not morally

[1] Williams himself refers to "empirical generalities of a kind which are the background to all problems of morality" (1973, 134), but it isn't clear that he really means his argument to be interpreted in the way I recommend below. Nonetheless, I propose that this is the most interesting objection in the vicinity, whether or not it is what Williams intended.

[2] This is the sense in which it is normal for cats to have four legs, even if every actually existing cat is an amputee (so that the median number of cat legs is three or fewer). This clarification secures the modal robustness of rational norms: It shouldn't turn out that the fundamental norms of rationality differ from world to world. But which psychological dispositions are statistically most *often* useful is something that radically differs from world to world, depending on external contingencies. So if we are to tie rationality to normal competence, we cannot mean 'normal' in the statistical sense. Rather, 'normality' in the relevant sense is something that is held fixed regardless of actual-world contingencies. One (controversial) way to precisify the notion is to assume that there is an *a priori* objective *probability distribution* over the possible worlds, and the 'normal' worlds are those that had the highest a priori probability of being actualized.

fitting. Add the following conceptual truth: if an agent that 'fits' some theory X is not yet *morally* fitting, then X is not the true moral theory. Now we can derive the critic's conclusion: Utilitarianism is a false moral theory.

In this chapter, I grant the first two premises for sake of argument. I seek to defend utilitarianism by instead rebutting the third premise: that a fitting utilitarian agent would be normally incompetent. Critics have thought this because they imagine the utilitarian agent as one who explicitly makes an expected utility calculation before each decision; who finds the needs of those before his eyes to be no more salient than those inaccessible and far away; and who is ready and willing to commit atrocities in the name of efficiency, without hesitation or regret. But, I will argue, these assumptions are mistaken. This is not what a fitting utilitarian agent would look like, as we can see when we take into account rational norms for fallible agents with human-sized minds.

Before we move on, I should say a little more to clarify the sense of 'fittingness' that features in this argument. This is a term of art that is meant to capture the intuitive idea of what's warranted or rational *from the perspective of* a moral theory — or, when used in a non-relational context, what's warranted or rational from the point of view of the *true* moral theory. It's fitting to desire that which is good or *desirable*, to admire the admirable, and so on. So, for example, if utilitarianism holds that what's good is just the welfare of sentient beings, then the fitting utilitarian agent is one who desires just the welfare of sentient beings. If such desires are shown to be not actually fitting, then that's just to say that utilitarianism is false: it makes mistaken claims about which things are desirable.

Of course, what's rational or fitting to desire may come apart from what it would be *best* or most fortunate to desire, just as what it's rational to believe (based on the evidence) may come apart from what it would be best or most fortunate to believe

(given various practical incentives).[3] So utilitarians can comfortably say that we 'ought', in the practical sense, to have whatever desires would be most fortunate, without thereby committing themselves to the view that those desires are *fitting*, or that their objects are truly *good*. And indeed, utilitarians have traditionally been much more interested in the question of which desires are best to have, than that of which are rational or fitting (according to their theory). But neglect of the latter question does not mean that there isn't a real question there to be asked. And, as we'll see, it's a question that the utilitarian *needs* to answer if they are to offer an adequate response to the argument laid out above.

3.1.2 Why the standard response fails

Consequentialists standardly distinguish between *criteria of rightness* and *decision procedures* (Bales 1971). Just because utilitarians hold that an act is right iff it maximizes expected utility (say), it doesn't follow that they recommend actually trying to calculate utilities in your everyday life. Indeed, given that such constant calculation would be predictably counterproductive (due to lack of time, misleading evidence, cognitive bias, setting bad precedents, etc. — see Mackie 1985), utilitarians would strongly recommend against it!

All this is true enough, but besides the point. The objection is not to utilitarianism's *recommendations*, but to its *implications*. There's a fact of the matter as to what the 'fitting' utilitarian psychology is, quite independently of what psychology utilitarianism *recommendeds* that we try to inculcate. But if the fitting-utilitarian psychology can be shown to be not actually morally fitting, that would — as previously explained — entail the falsity of utilitarianism as a moral theory.

Any moral theory has some implications for what kind of psychology is 'fitting' or rational from a moral point of view. At a minimum, as we've seen, one's theory of the

[3] For more on this distinction, see my second chapter, 'The Fitting and the Fortunate'.

good commits one to holding certain objects (namely, the things identified as good) to be *fitting to desire*. Again, that's not to say that a utilitarian must recommend that we try to acquire fitting desires — it's an open empirical question whether being 'rational' in this way would promote utility in actual circumstances. So that's not the issue. Rather, the issue at hand is just whether the kind of mindset utilitarianism *implies* is rationally fitting *really is* so. If it's not, then the theory is shown to be false, in virtue of having false implications. For example, if utilitarianism implied that people are 'replaceable', in the sense that it's rationally fitting to desire each person's welfare merely as a means to promoting the aggregate welfare, then (assuming we're right to doubt the latter claim) that would be grounds for thinking that the theory must be mistaken.[4]

So we can't just ignore decision-procedures and other psychological elements. And nor can we merely settle for identifying those which are most conducive to utility, and thus *recommended* by utilitarianism. As normative theorists, interested in whether or not utilitarianism is a true moral theory, we must also investigate what kind of mindset would be a *rational* mindset, were utilitarianism true. We can then test whether this fitting utilitarian mindset meets the minimal requirements for rationality, such as the 'normal competence' test proposed in §3.1.1, and hence whether utilitarianism itself remains an eligible moral theory.

In the following sections, I explore two very different strategies for constructing a non-defective utilitarian psychology in answer to this challenge. §3.2 explores the Railtonian "sophisticated" psychology, with non-utilitarian desires. §3.4 sets out my preferred "subjective" account, arguing that critics are mistaken to assume that a fitting agent with utilitarian motivations would be guided by explicit "expected utility" calculations.

[4] Happily, as argued in my second chapter, 'The Fitting and the Fortunate', utilitarianism does *not* imply such replaceability!

3.2 Sophisticated Utilitarianism

3.2.1 Explication

Railton (1984) contrasts two kinds of hedonistic (or, more broadly, consequentialist) psychologies: 'subjective' and 'sophisticated'.[5] The subjective hedonist is solely motivated by concern for his own happiness. However, the 'paradox of hedonism' suggests that such a person is likely to end up quite unhappy. Happiness may be better achieved by those who are motivated by other concerns. Railton thus introduces the sophisticated hedonist — let's call her 'Sophie' — who "aims to lead an objectively hedonistic life (that is, the happiest life available to [her] in the circumstances) and yet is not committed to subjective hedonism." Sophie may thus possess and act on distinctively non-hedonistic motives — e.g., concern for others — if such desires are conducive to her living a happier life overall.

Once she has moved beyond subjective hedonism, and acquired a happy collection of non-hedonistic motivations, we may begin to wonder in what sense Sophie is still a "hedonist" at all, rather than a whole-hearted pluralist. What sets Sophie apart, according to Railton, is that her psychology continues to be regulated by a **counterfactual condition** according to which, despite her various desires, she "would not act as [s]he does if it were not compatible with [her] leading an objectively hedonistic life."

This requires some unpacking. One might be tempted to interpret the counterfactual condition as applying at the level of individual acts, so that Sophie always chooses the hedonistically recommended action. But this would seem to require an overriding desire for happiness. The most that can be said for Sophie, on this interpretation, is that her non-hedonistic desire may be the one that is causally operative in cases of

[5] Below I follow Railton in focusing on hedonism for simplicity. But everything I say should carry over, in obvious fashion, to the case of utilitarianism.

motivational overdetermination.[6] But her hedonistic desire is always there, waiting in the wings, ready to swoop in and exert its overriding force whenever it appears that she is about to make something other than the hedonistically-recommended choice. Sophie thus looks little different from the subjective hedonist, on this interpretation.

Alternatively, we may interpret the counterfactual condition as applying at a more global level (as suggested by Railton's reference to the "objectively hedonistic *life*"). Whereas the subjective hedonist regulates her individual actions according to hedonistic norms, Sophie's hedonism instead regulates her desires and dispositions. So, for example, her pro-friendship disposition may lead Sophie to perform individual acts that reduce her happiness — e.g. answering her friend's distraught 3 a.m. call — but her 'hedonic monitor' is not triggered to intervene unless it becomes clear that the relationship *as a whole* is detrimental to her happiness, such that she would be better off in the long run with different desires and dispositions.

Some questions remain concerning the precise details of how this regulative mechanism is supposed to work. In particular, we may wonder whether Sophie has an overriding desire to *possess hedonically fortunate dispositions*, that she will act upon (overriding her other, non-hedonistic motivations) whenever she's in a position to do so. This may still seem too close in structure to the psychology of the simple consequentialist. It so happens that Sophie's overriding desire, due to its quirky content, applies to fewer situations than does the corresponding desire of the subjective hedonist. (More of our choices potentially impact our happiness than potentially impact whether we have happiness-promoting dispositions.) But this looks more like a difference in degree than a difference in kind. Indeed, in light of the structural parity, such an agent may be better described as a straightforward subjective maximizer of happiness-promoting dispositions, rather than a sophisticated maximizer of happiness!

[6] I owe this suggested interpretation to Michael Smith.

We may do better to interpret Sophie's hedonism as manifested not in a *desire* at all, but rather a higher-order mechanism that serves to regulate her desires through some sub-personal causal process. The key difference is that this hedonistic mechanism, unlike a desire, never directly manifests itself in action. It is not *itself* a motivation that she may act on (though it may cause her to acquire some independently motivating hedonistic desires, insofar as these would cause her to live a happier life). Its control over her actions is instead wholly indirect: The hedonic monitor shapes Sophie's desires in hedonically fortunate ways, and then she acts on *these desires*, whatever they may be.

One worry for this view is that the hedonistic desire-regulating faculty begins to look like an 'external' imposition on Sophie — turning her into a kind of puppet. To avoid this problem, we need to ensure that the regulating faculty is in some sense under Sophie's rational control. It is *because* Sophie believes in (or has a deep-rooted standing commitment to) hedonism that the faculty regulates her desires according to hedonistic goals. If she came to fully believe and endorse some other normative theory, say utilitarianism, then the faculty would regulate her desires according to utilitarian goals instead. The 'sophisticated' psychology is thus best described in two parts: First, there is the agent's overarching "primary goal", which she may identify with during reflective moments, but which does not tend to directly motivate her actions. Instead, she is moved by the "secondary" desires and dispositions that are produced and regulated by a mechanism that is responsive to her primary goal.

3.2.2 Evaluation

Supposing that the psychology described in §3.2.1 is coherent, it's an interesting question how exactly we should evaluate it. Suppose that (egostic) hedonism is true, so that one's own pleasure is the only end that's truly desirable, or worth pursuing. Sophie then seems irrational, in that her desires do not conform to the normative

facts about what's desirable, and her actions likewise fail to be sensitive to hedonistic reasons: She often benefits others at her own expense. On the other hand, she is not *completely* insensitive to reasons: Her desire-regulating faculty ensures that she maintains the desires that (the evidence suggests) it is best for her to have — and if circumstances change, so will her dispositions. This suggests an important sense in which Sophie's reflective hedonism is ultimately 'in control', even if it is not what moves her. We may thus need to draw a distinction between (local) act and (global) agent rationality, allowing us to say that *Sophie* is rationally fitting or responsive to reasons, even if her particular *actions* are not.

It's worth noting that even this vestige of rational sensitivity may, in special circumstances, make her worse off. Consider Parfit (1984)'s example of the society of perfectly rational egoists, some of whom come to realize that it will advance their interests to become irrational in a specific respect: namely, if they become transparently disposed to follow through on their threats regardless of the costs to themselves. Such a "threat-fulfiller" can then strap a bomb to his chest, and threaten an egoist that he will detonate it (killing them both) unless the egoist complies with his whims. He can safely make such threats, because he knows the egoist would sooner comply than die. As Parfit further shows, the rational response for the remaining egoists is to turn themselves into transparent "threat ignorers", who are stably disposed to (irrationally) ignore threats no matter the costs to themselves. A threat-fulfiller will leave the ignorers alone, because he knows that if he were to threaten them, they would ignore him, and he would then detonate the bomb, killing them both. (Note that the threat-fulfiller will not *issue* threats that he expects will make him worse off. It is merely *fulfilling* threats that he does blindly.)

In comparison to the pure threat-ignorers, Sophie is more apt to have her rationality exploited. Given transparency, the threat-fulfiller will know that if he threatens Sophie, she will comply. For Sophie's regulating mechanisms will not allow her to

maintain a disposition once it becomes clear that it is disastrous for her long-term happiness. And a threat-ignoring disposition becomes clearly disastrous as soon one is actually issued with a credible apocalyptic threat. So, a threat-fulfiller will know that he can safely threaten Sophie, and she will (if necessary change her dispositions and) comply rather than die. To avoid such exploitation, Sophie would have to alter her psychology so that she would become a *pure* (unregulated, insensitive) threat-ignorer — at which point she would no longer be a sophisticated hedonist. She would just be (however fortunately) plain irrational.

We thus find that a Railtonian sophisticated hedonist (or, more broadly, sophisticated consequentialist) psychology is by no means guaranteed to endorse itself as the most fortunate psychology to possess in every possible situation. But it offers a suggestive alternative to the standard conception of a rational psychology. Insofar as we are drawn to the idea that rationality should not *normally* be a curse (even if it may be in certain special circumstances), we may see Sophie's two-level psychology — with its capacity for her primary goal to control and regulate her secondary, action-guiding motivations — as an improvement over the subjective hedonist's unitary motivational structure. While acknowledging that Sophie's actions are often locally irrational (by hedonist lights), we may be more concerned to evaluate her global rationality as an agent. In this respect, at least, she may at first glance seem more reasonable.

I think there are important grounds for doubting this conclusion, however. Let's return our attention from hedonism to utilitarianism. The sophisticated utilitarian — call her 'Sophu' — will have whatever motivations are most conducive to promoting the general welfare. So, in particular, if an evil demon threatens to torture an innocent population unless Sophu comes to intrinsically *want*[7] them to suffer, then Sophu will

[7] One might wonder if Sophu's resulting pro-suffering desire will be merely instrumental, since it is produced (by her regulating mechanism) as a means to promoting welfare. But we must take care to distinguish conditions on the desire's *existence* from conditions on its *content* (or, in other words, to distinguish a desire's persistence conditions from its fulfilment conditions), and I take the instrumental/intrinsic distinction to concern the latter. It may be that Sophu's regulating mechanism ensures that the desire's existence is *contingent* upon its promoting utility, but that doesn't mean

be led to acquire this fortunate but malicious motivation.[8] This is a good outcome, in the circumstances, as it prevents a lot of suffering. But if any desire is held to be irrational or unfitting by utilitarian lights, it is surely an intrinsic desire that others suffer. Sophu has, quite virtuously, made herself vicious by utilitarian lights. And note that it is not just her *actions*, but her *desires* — her very *self*, we might think — that is impugned here.

The advocate of sophisticated utilitarianism might at this point defend Sophu by pointing out that her *deepest commitments* remain pure and altruistic, even as they respond to the unfortunate circumstances by shaping her motivations in this malicious-but-instrumentally-valuable direction. So there at least remains *something* fitting about Sophu's psychology. The tricky question, which I will not resolve here, is whether this is enough to allow her to qualify as a fitting utilitarian agent *overall*.

3.3 Rational Transmission

It seems plausible to think that there's a tight connection between (i) the rationality of acquiring and maintaining a disposition, and (ii) the rationality of 'acting on' the disposition, i.e. performing an action that the disposition characteristically disposes you towards. One candidate connection is suggested in the following simple principle of rational transmission:

> **(RT-past)** For any disposition D and act A that is characteristic of D: *If it was rational to acquire D then it is rational to perform A.*

that considerations of utility enter into the *content* of the desire. What Sophu ends up wanting — i.e., the content or object of her desire — is simply *that people suffer*, not some other end to which this suffering is a means. She becomes motivated to pursue suffering for its own sake. It's just that her motivations will change when they cease to promote (expected) utility.

[8] At least, this is so on my interpretation of the 'sophisticated' psychology. Mason (1999, 256) suggests an alternative view on which "We can develop new motives from old motives, but only when they are consistent." This would rule out a sophisticated utilitarian acquiring malicious motivations, however beneficial they might be. However, it's not clear to me whether this is meant to be a conceptual constraint on rational agency, or just a contingent empirical hypothesis about how new motivations actually develop in people.

But Parfit (1984)'s above-described case of the threat-fulfillers casts doubt on this principle. It may well be rational for a self-interested agent to acquire the threat-fulfilling disposition, but if (through some irrational quirk) a threatened target unexpectedly ignores the agent's apocalyptic threat, it is surely *not* rational for the agent to follow through and blow themselves up. Such disastrous stubbornness would seem, on the contrary, quite crazy.

Gauthier (1997) is not wholly convinced by this counterexample to RT-past, but suggests and endorses two weaker transmission principles, which we may formulate as follows for any disposition D and act A that is characteristic of D:

> **(RT-timeless)** If it was rational to acquire D *and is better to maintain D than never to have possessed it at all*, then it is rational to perform A.
>
> **(RT-present)** If it was rational to acquire D *and is rational to maintain it presently*, then it is rational to perform A.

However, Parfit (2011, Appendix B) points out that even these weakened transmission principles are susceptible to counterexamples, such as:

> *Schelling's Case.* A robber threatens that, unless I unlock my safe and give him all my money, he will start to kill my children. It would be irrational for me to ignore this robber's threat. But even if I gave in to his threat, there is a risk that he will kill us all, to reduce his chance of being caught. [... I]t would be rational for me to take a drug that would make me very irrational. The robber would then see that it was pointless to threaten me; and since he could not commit his crime, and I would not be capable of calling the police, he would also be less likely to kill either me or my children. [...] But while I am in my drug-induced state, and before the robber leaves, I act in damaging and self-defeating ways. I beat

> my children because I love them. I burn my manuscripts because I want to preserve them.

Parfit stipulates that these destructive acts are not necessary to convince the robber that you are irrational. So they have no good effects, though they stem from a disposition that it is worthwhile (for extrinsic reasons) to possess. Are these acts rational? I share Parfit's sense that they are not. So all of the transmission principles surveyed thus far fail.

The fundamental explanation for this disconnect is that an agent's dispositions can have other consequences besides producing downstream acts in the agent herself. In particular, you might be harmed or rewarded directly on the basis of whether you possess some disposition, independently of whether you act on it. This suggests that we can distinguish (i) dispositions that have high expected value, all things considered, and (ii) dispositions that have high expected value *in respect of the downstream actions they'll tend to produce*. We can call the former class of dispositions 'desirable', and the latter 'well-calibrated'. Dispositions that are desirable but *not* well-calibrated we may call 'extrinsically desirable'. It is these extrinsically desirable dispositions that feature in Parfit's cases of 'rational irrationality', i.e. whereby it is rational to acquire and maintain such a disposition, but not to act upon it.[9]

While acknowledging this possibility, we may still think that there must be *some* 'transmission' principles according to which the rational status of a general rule or disposition can be inherited by the particular acts it prescribes. And, indeed, the distinction I've just highlighted suggests an obvious candidate principle: we just need to restrict the dispositions in question to those that are 'well-calibrated', i.e. desirable for their (expected) impact on your downstream actions, rather than for extrinsic reasons. Consider the following transmission principle:

[9] This parallels the familiar distinction between 'object-given' and 'state-given' reasons. The dispositional state is a useful state to be in, but the (metaphorical) 'object' of the state — the set of actions it disposes you towards — does not merit such a disposition.

> **(RT-Calibrated)** For any dispositional set D and act A that is characteristic of D: *If D is* well-calibrated*, i.e. expectably good to possess in virtue of the downstream actions it tends to produce, then it is rational to perform A.*

Alternatively, we may formulate the principle in terms of rules rather than dispositions:

> **(RT-CalibratedRule)** If S rightly adopts a rule R as maximally well-calibrated (i.e. expectably better to internalize than any conflicting rule, in virtue of the downstream actions it prescribes) and R prescribes ϕ-ing in circumstance C, then when S is in circumstance C, S rationally ought to ϕ.

Such a principle may be supported by considerations of meta-coherence. Ex hypothesi, the rule R offers the most reliable guidance available to S — in particular, it produces good actions more reliably than does attempting to autonomously determine what the best result would be in each case. (And it is also more reliable than any identifiable alternative, e.g. "following R except in circumstances with the subjectively distinguishing feature F.") So, any given departure from R can be expected to have worse results than would be obtained by following R. So in any given case, the agent should follow R's advice.

Actually, this isn't quite right. The argument from metacoherence only supports acting on rules that improve the agent's actions due to serving as an *epistemic guide*, like the rule against killing people even when murder might (prima facie) *seem* to promote utility. But in non-ideal agents, a disposition might also conduce to good actions for a very different kind of reason, as I will illustrate below.

Meet Cam, a callous consequentialist. Cam is one of those utilitarians who likes humanity but not people so much.[10] Due to his lack of regard for those around

[10] Just to clarify: This nasty streak is not, of course, any part of the fitting utilitarian psychology!

him, he tends to act insensitively, and makes other people (not least his poor family) miserable. Upon reflection, Cam recognizes this to be unfortunate. But he lacks the strength of will to reliably act better in such situations. So he takes a pill that makes him a much more caring and loving person. He is now disposed to attend disproportionately to the welfare of those that are close to him. This causes him to act in much better ways: in particular, he finds it easier to refrain from making the kinds of insensitive remarks that previously caused so much harm. The one downside is that he is now much less inclined than before to donate to GiveWell-recommended charities that promote the impartial good. He would rather spend that money on his family. This is harmful, but (let's suppose, perhaps unrealistically) not nearly as harmful as Cam's callous actions had been.

The structure of the case is that Cam was previously *weak-willed* in a very bad way. He then acquired a disposition that helped him to overcome this weak will, and so perform better actions — though at the cost of acquiring a new (less bad) weakness. Because it is clear (by stipulation) which of his newly-disposed actions are better than before and which are worse, the epistemic argument for following a generally beneficial disposition no longer applies. (The overall benefits of the change aren't evidence that his giving less to charity, in particular, is beneficial. On the contrary: *this* action remains clearly bad, it's just that value of his *other* actions outweighs this localized badness.) So, we find that it's right for Cam to acquire this disposition in virtue of the actions it conduces to in general, but some particular actions it conduces to may still be considered wrong, or less than perfectly rational from a utilitarian perspective. We can thus have cases of what Parfit (1984) would call 'blameless wrongdoing', without having to appeal to dispositions that have good effects besides action. (Of course, the disposition may also be good for other reasons — the point is just that my case doesn't rely on this.)

Perhaps we can get around such cases by restricting the transmission principle to dispositions that are *maximally* well-calibrated:

> **(RT-CalibratedMax)** For any dispositional set D and act A that is characteristic of D: *If D is* maximally *well-calibrated, i.e. the expectably best dispositional set to possess in virtue of the downstream actions it tends to produce, then it is rational to perform A.*

This assumes that there is a possible dispositional set that achieves the best of both worlds, relieving Cam of his previous character flaws without adding any new ones. But we may doubt whether that is always possible, and it also restricts the usefulness of the principle for non-ideal agents. Perhaps the best we can say is that when a rule or disposition is well-calibrated for action *in virtue of serving as an epistemic guide*, then following the guidance of the rule or disposition is rational.

A deeper cause for hesitation comes from considering cases where you adopt a rule as a hedge against (as it happens, misleading) evidence that you might be biased in your subsequent judgments. For example, the rule might tell you to disregard certain first-order evidence, because you can't be trusted to evaluate it rationally. But if you actually *are* capable of evaluating it rationally, then we may think that there's an important sense in which you rationally ought to be guided by the (first order) evidence. Or, even if the higher-order evidence makes some contribution, we may still doubt the radical claim that it *completely swamps* the first-order evidence in determining what you rationally ought to do (cf. Kelly 2010). In that case we may similarly doubt that there are any true and interesting (non-trivial) principles of rational transmission from dispositions to individual acts.

In sum: While I am uncertain that any such principle is ultimately vindicated, focusing on the subset of dispositions that are *well-calibrated*, in my described sense, would seem to give us the best shot. And, as we will see, these are just the dispositions that may be possessed by the "subjective" act utilitarian agent, in contrast to the

unfitting but (extrinsically) desirable dispositions that we saw could be part of the "sophisticated" utilitarian psychology.

3.4 The Act Utilitarian Agent

Suppose we accept my earlier suspicion that "sophisticated" psychologies, with their extrinsically desirable dispositions, are not rationally fitting psychologies. The remaining option for defending utilitarianism against the argument of §3.1.1, is to spell out a non-defective "subjective" utilitarian psychology. In this section, I will attempt this task, by drawing on the critic's assumption that there are rational norms for human-sized minds that render an agent competent to act in normal circumstances. I will especially make use of the idea that our account of the fitting utilitarian agent, while restricted to utilitarian motivations, may at least appeal to *well-calibrated*, if not merely extrinsically desirable, guiding dispositions. This restriction is one of the main features that sets apart my straightforward account of the fitting utilitarian psychology from the "sophisticated" view explored in §3.2.

3.4.1 Motivating vs. Guiding Dispositions

Let's begin by distinguishing what I'll call 'guiding' and 'motivating' dispositions.[11] Our non-instrumental desires or *motivations* are our driving concerns, or what move us to action. They represent the goals we hope to realize through acting. On the other hand, this motivational 'oomph' can be steered or *guided* by strategies and heuristic dispositions that shape our behavioural responses in pursuit of those goals. We may think of our guiding dispositions as, roughly, the psychological manifestation of instrumental rationality. They take our desires as inputs, and output a suitable action or intention.[12]

[11] Thanks to Philip Pettit for his assistance in formulating this distinction.

[12] I remain neutral on the question of whether practical reasoning is best understood as concluding in action or intention.

The standard caricature of the utilitarian agent assumes that we can "read off" both kinds of dispositions from the moral theory. From its theory of the good — the view that what matters is just the welfare of sentient beings — we get the fitting utilitarian motivations. That much I agree with: the fitting utilitarian will desire the welfare of sentient beings.[13] But the standard caricature also takes the 'maximizing' aspect of utilitarianism to settle the *guiding* dispositions of the fitting utilitarian agent: they will (allegedly) decide how to act by, in each instance, conducting an expected-utility calculation, and then perform whatever action they judge to have the highest expected utility. It is this feature of the imagined utilitarian agent that is responsible for so much of their apparent defectiveness. And it is this feature that I deny we should attribute to the fitting utilitarian agent.

Instead, I propose that our choice of moral theory only commits us to the fittingness of the corresponding *motivating* dispositions. When it comes to the fittingness of guiding dispositions, this is instead determined by our independent — morally neutral — account of instrumental rationality. Drawing from our critic's understanding of *rationality as normal competence*, we can elucidate the fitting guiding dispositions as those that are prerequisites for normally competent agency in creatures with human-sized minds.

My strategy for responding to the self-effacingness objection of §3.1.1 is thus as follows. I will offer a brief sketch of some 'well-calibrated' guiding dispositions which I take to be (a) prerequisites for normally competent human-like agency, and hence (b) rationally fitting for agents with human-sized minds. I will then show how an agent with fitting utilitarian motivations could also possess these well-calibrated guiding dispositions. The result is a non-defective, normally competent, fitting utilitarian

[13] Though it's an important question whether we interpret this as a single monolithic desire for aggregate welfare, or — as I prefer — a plurality of desires, one for *each* sentient being's welfare. See my second dissertation chapter, 'The Fitting and the Fortunate'.

agent. I will wrap up by illustrating how my 'well-calibrated' fitting utilitarian can be used to address several prominent character-based objections to utilitarianism.

3.4.2 Defective Deliberation and the Well-Calibrated Agent

Let me spell out four central features of the fitting agent's guiding dispositions. Firstly — as perhaps the most obvious prerequisite for competent agency — we have *epistemic rationality*: that is, the agent must have well-calibrated expectations about their environment. They cannot take the roar of a dangerous predator as evidence that a cute puppy awaits them outside. They need to have generally reasonable beliefs about their environment, and about what would be effective means for realizing their ends (whatever those might be).

Next, at the borderline of the epistemic and the practical, we will find constraints on how the agent is disposed to allocate their limited *attentional* resources. They must be generally attentive to possible threats and opportunities in their immediate environment, while also — in a calm moment, when appropriate — considering more abstract mental models of past and possible future scenarios (for sake of planning, self-evaluation, etc.). The details aren't too crucial for my purposes, but as we'll see, it's important that the fitting agent not dwell excessively on the past.

Thirdly, the competent agent requires well-calibrated habits, instincts, or subpersonal "predispositions" (Pettit and Brennan 1986) — an "auto-pilot" set, e.g., to avoid pain, be cooperative, and help others in need — to secure effective automatic behaviour in normal circumstances.One reason for this is that in time-critical situations, the agent cannot afford to pause to reflect on their situation at all. Often, a competent agent will be moved immediately (without conscious deliberation) to act, upon registering pertinent information about their environment. This is no mere behavioural reflex, as the agent is genuinely acting for reasons. But the rational processing goes on 'below the surface', as it were.

Once equipped with such well-calibrated predispositions, the fitting agent may act on them without need for excessive self-monitoring or executive control, and — in so doing — they may trust that they are acting for the best. Our fitting agent may, in this way, reap the practical benefits of 'satisficing' without the theoretical baggage.[14]

The fourth and final component that I'll discuss here is the possession of well-calibrated *triggers* for executive oversight. On pain of regress, we cannot deliberate about whether to deliberate. So, as previously noted, the agent's *default* guidance must be from non-deliberative "predispositions". But when these are not up to the task — when, say, the agent is faced with novel or complex circumstances for which their predispositions aren't so well calibrated to deal with — the agent's sub-personal mechanisms must recognize this and respond by triggering explicit deliberation on the part of the agent.

In summary: the fitting human-like agent — if they are to be capable of acting competently in a wide range of 'normal' circumstances — will rely heavily on well-calibrated predispositions, rather than explicit deliberation or calculation, to guide their actions in pursuit of whatever their goals may be. And this will be so even if their goal is to promote the well-being of sentient creatures as much as they are able. This, I propose, is how we should understand the fitting utilitarian agent. They have straightforwardly utilitarian desires, which are then translated into action via the above-described 'well-calibrated' guiding dispositions.

3.4.3 Addressing The Objections

We are now in a position to assess how my conception of the fitting utilitarian agent stands up to various anti-utilitarian objections.

[14] Cf. Slote and Pettit (1984). Slote's satisficing consequentialist merely aims to achieve 'good enough' consequences, which makes sense as a practical strategy but seems rather more puzzling as an account of the rationally warranted *ultimate aims*.

We can first note that my 'well-calibrated' fitting utilitarian will not be "constantly calculating". Absent any triggering of their executive faculty, the fitting utilitarian will respond directly to the salient needs of others — a child drowning in a pond, say — without mediation by explicit deliberation, let alone abstract judgments of "permissibility". In this way, the fitting utilitarian will not exhibit what Williams (1982) famously called "one thought too many".[15]

The fitting utilitarian's reliance on generally-reliable predispositions also undermines the objection that they would engage in "marginally-beneficial rule-breaking", such as breaking a promise whenever the benefits from doing so seem to even slightly outweigh the costs.[16]

Because overt calculation often goes awry, the competent utilitarian will — as we've seen — rely heavily on her generally reliable predispositions in everyday life, only pausing to reflect when her well-calibrated sub-personal mechanisms alert her to the need (say due to complex novel circumstances, that her "auto-pilot" wasn't designed to deal with). Everyday promise-keeping is not exactly novel, so for the fitting agent the question whether to keep a promise *shouldn't even arise*, unless there's something special about the situation that calls for her executive oversight.

That's enough to defeat the claim that the fitting utilitarian would commonly engage in marginally-beneficial rule-breaking. But we may draw an even stronger conclusion. For suppose that our agent's executive oversight happens to be triggered. In a typical case, what should she conclude? We can stipulate that in fact the outcome would be marginally better if she broke her promise, but presumably the agent herself

[15] I should mention that the traditional worry here is not just that the extra thought will make the agent too slow to act, but that it reveals something distressingly 'alienated' about his psychology, and the seemingly 'instrumental' nature of his concern for others. I further address such concerns in chapter 2.

[16] Hooker (2000) raises the objection in terms of act utilitarianism implying that it's *objectively right* to break the rules if the benefits of doing so actually marginally outweigh the costs. I don't have particularly strong intuitions about objective rightness, as opposed to *what a moral agent would do*, so I restate the objection here in the latter terms. (Note that in chapter 2, the intuitions I appeal to primarily concern objective *goodness*, not objective rightness.)

will not have any easy way of knowing this. (Among other things, she'd need to first consider the possibility of self-serving bias corrupting her judgment, and also to weigh the apparent benefits of rule-breaking in this instance against the long-run value of retaining a reputation for trustworthiness.) Maybe if she heard the booming voice of God reassuring her of this fact, then she could rationally go ahead and break her promise without further worry. But in *ordinary* circumstances — as we're supposed to be concerned with here — it's almost never going to be *clear* that rule-breaking is beneficial unless it is significantly (not merely marginally) so.[17]

So our agent is faced with an immediate choice: she can (i) break the rule even though it's not yet clear to her whether this would have good results on net; (ii) sink further cognitive resources into investigating a question that she probably shouldn't have bothered to ask in the first place; or (iii) simply keep her promise and turn her attention to more important matters. It seems pretty clear that, in this sort of case, option (iii) is the way to go.[18]

In sum: Breaking a rule will generally only be obviously worthwhile in cases where it is also of significant benefit (in which case many would approve of rule-breaking anyway). If it's only of marginal benefit, this fact typically won't be sufficiently clear for a reasonably self-doubting, fallible agent to immediately act upon. And the low potential payoff means that it isn't really worth inquiring further: better just to stick with the generally-reliable rule of thumb. So a rational utilitarian generally won't be found engaging in marginally beneficial rule-breaking after all. (They'd even share our intuition that there's something awfully dubious about any agent who would

[17] Moore (1903a) claimed that agents will never be in an epistemic position to justfiably break such rules, but we needn't be quite so pessimistic.

[18] In special circumstances, option (ii) may be truly costless, and so there's a possibility that the agent could reasonably undertake such an investigation, and responsibly reach the true conclusion that breaking the rule really is justified in this case. But this won't be typical, and the crucial point for my purposes is just that one's *prima facie* utility judgments won't provide sufficient justification for breaking generally-reliable rules.

act that way.) This gives them the kind of stable predictability needed for others to regard them as eligible and (more or less) trustworthy partners for social cooperation.

This discussion brings out the fact that the standard caricature of a utilitarian agent assumes that they will be unreasonably overconfident in their ability to calculate utilities accurately. But even if a utilitarian *initially* judges (just based on the first order evidence) that they would do best to break some generally beneficial rule, they may also realize that most people who make such judgments in similar situations are mistaken. Since they have no particular reason to think that they are one of the lucky few who make this judgment correctly, the general fact serves as a kind of higher-order evidence that their initial judgment was mistaken. All things considered, then, a reasonable expected utility judgment should, in this sort of circumstance, end up reinforcing the general rule rather than licensing typically-misguided unilateral rule-breaking.

The objections considered thus far — that the utilitarian would have "one thought too many", and that they would engage in "marginally-beneficial rule-breaking" — suggest the need to distinguish (i) the appropriate answer to a question, and (ii) whether a well-functioning agent would ask that question in the first place. The need for this distinction becomes especially apparent when we consider the following objection from Stocker (1989, 321):

> Maximizers hold that the absence of any attainable good is, as such, bad, and that a life that lacks such a good is therefore lacking. I disagree. One central reason for my disagreement stems from the moral psychological import of regretting the absence or lack of any and every attainable good. This regret is a central characterizing feature of narcissistic, grandiose, and other defective selves. It is also characteristic of those who are too hard on themselves, who are too driven and too perfectionistic.

This objection seems to me misguided. It may be unfortunate, and indeed even inappropriate ("defective"), to actively regret every little regrettable thing. But those things may be regrettable all the same. Crucially, this is not to say that a rational agent must regret them. It is more like a hypothetical imperative: *if* you closely attend to the features in question, this should induce in you feelings of regret. But it may be a kind of rational defect to attend to the wrong things, if there are more pressing matters to attend to. As we saw in §3.4.2, the fitting agent would allocate their attentional resources in a way that avoids excessive dwelling on hypotheticals. So we can agree with Stocker that the agents he describes are defective, without thinking that the maximizing utilitarian would exhibit any such trait. On my picture, the utilitarian will have only a conditional disposition to regret the lack of a good *insofar as she attends to this lack*. But she'll usually have more important things to attend to, so she shouldn't actually end up actively regretting things very often at all. She is, in this sense, appropriately respons*ive* to reasons for regret, without having to be constantly respond*ing* to them.

I've now shown how the well-calibrated fitting utilitarian avoids three of the 'character-based objections' extant in the literature. Equipping the utilitarian agent with well-calibrated guiding dispositions helps to undermine claims that the fitting utilitarian psychology is so typically self-effacing (across a wide range of normal circumstances) as to warrant charges of intrinsic defectiveness. In the course of this defense, I have tried to employ a fairly conservative methodology: Beginning from assumptions about rationality that underpin the original objection, I have drawn out a conception of a *normally competent* fitting utilitarian agent, thus rebutting the charge that the fitting utilitarian is too incompetent to qualify as a truly morally fitting agent.

3.4.4 Act vs. Rule Consequentialist Agents

In light of my appeal to rules and dispositions, some readers may be puzzled by my labelling the resulting agent a fitting 'act utilitarian' agent, rather than a rule utilitarian one. To avoid any confusion on this front, let me wrap up by briefly characterizing what I take to be the two main differences between (fitting) act and rule utilitarian psychologies.

Firstly, while both make use of rules, they do so in very different ways. The act utiliarian adopts 'rules of thumb' for *instrumental* purposes, but their fundamental aim (in acting) makes no essential reference to rules: they just want to bring about the best possible outcome, and refraining from deliberation is one strategy they might employ, at appropriate times, as a means to this end. Rule Consequentialism, by contrast, builds reference to rules into its criterion of right action, and hence the corresponding 'fitting psychology' must likewise accord some fundamental, non-instrumental significance to rules. (This then opens them up to distinctively characterological objections of 'rule-worship'.)

A second, more straightforward difference is that they may employ rules with very different contents. I've suggested that a fitting act utilitarian could (whilst retaining their fitting character) only make use of 'well-calibrated' dispositions. But insofar as rule consequentialism appeals to rules that it would be good to internalize *for whatever reason*, they may well end up calling 'fitting' even dispositions that are merely extrinsically desirable. In other words, the fitting rule consequentialist agent would look much more like the kind of 'sophisticated' agent described in §3.2. Insofar as we doubt that such a psychology *really is* morally or rationally fitting, this could provide the basis for a new argument against rule consequentialism — though not one that I have space to develop here.

Chapter 4

Virtue and Salience

This chapter explores two ways that fittingness assessments of a (cognitively limited) agent's character are properly influenced by what the agent finds salient or attention-grabbing. First, I argue that ignoring salient needs reveals a greater deficit of benevolent motivation in the agent, and hence, on the 'Quality of Will' account, renders them more blameworthy. I use this fact to help explain our ordinary intuition that failing to give to famine relief (for example) is in some sense *less bad* than failing to help a drowning child right before your eyes, in a way that's compatible with the consequentialist's contention that there's no principled reason to see the one life-saving act as any more or less choiceworthy than the other. Second, I argue that alleged 'virtues of ignorance' (e.g., modesty, believing better of friends than the evidence supports, etc.) are better understood as 'virtues of salience'. The modest person, for example, needn't have any *false beliefs* about their own accomplishments; what sets them apart from the immodest is instead that their own accomplishments aren't as *salient* in their thoughts — their attention is not constantly directed back towards themselves in the manner of the immodest. Virtue may thus make demands upon what we find salient.

4.1 Salience and Quality of Will

4.1.1 The Puzzle

Singer (1972) invites us to imagine a child drowning in a pond, whom we could save at the cost of ruining our expensive clothes in the muddy water. It seems clear that we ought to save the child, no matter the (comparatively insignificant) financial cost to ourselves. This then motivates what we can call Singer's *Insignificant Sacrifice Principle* (p.231): "[I]f it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it." However, as Singer goes on to point out, we are constantly violating this principle by failing to donate as much as we could to effective charities that address various serious-and-yet-easily-preventable harms caused by global poverty: malnutrition, lack of basic medical care, vaccinations, etc.[1] What should we think of this practical inconsistency?

When we reflect on what's at stake, I find it quite plausible that we really should prevent these grave harms rather than buying unnecessary luxuries for ourselves. (Call this *the Act Evaluation*.) That much of Singer's argument seems right.[2] But the analogy between the drowning child and the global poor may also be taken to suggest a much more troubling conclusion, via the following argument:

1. It would be morally monstrous to do nothing and let the drowning child die.

2. Saving a distant stranger's life by donating to an effective charity is relevantly similar to saving a nearby drowning child.

3. So, it would be morally monstrous to let a distant stranger die when we could have saved their life by donating to an effective charity.

[1] For research into which charities are in fact the most effective, see `www.givewell.org`.
[2] I won't be defending this claim any further here, but see, e.g., Unger (1996), Kagan (1991), and Cullity (2006).

Here the conclusion is not just about the status of the *act* of helping — that it merits choosing, or ought to be done — but about the moral *character* of the agent who fails to act as they ought in this case. It is, according to this argument, no minor failure, but one that renders us *morally monstrous* or blameworthy to the highest degree. And this claim — call it *the Character Evaluation* — seems entirely incredible. It just isn't plausible that in failing to save distant strangers we reveal our moral character to be as bad as someone who callously watches a child drown and does nothing about it.

4.1.2 The Solution

Appeal to a *Quality of Will* account of blameworthiness can help to resolve this puzzle, showing how we can accept Singer's Act Evaluation without committing ourselves to the implausible Character Evaluation. According to Quality of Will accounts, an agent is blameworthy to the extent that their actions manifest an insufficient degree of good will towards others (Strawson 1962; Arpaly 2003). An agent may be understood as having a "sufficient degree of good will" when their desires for others' welfare are sufficiently weighty in relation to their personal or self-interested desires. The less that I care about others, and hence the more harms that I'm willing to impose on them for the sake of my own lesser benefit, the morally worse my character becomes. When I perform actions that manifest this lack of concern for others, I am blameworthy in proportion to the moral inadequacy of the desires that I thereby act upon. (Someone who acts with at least some minimal concern for others will be less blameworthy than someone who acts in complete disregard of others' interests.)

So that's the basic picture. But now notice the following fact about human psychology: Our actions are determined not only by the strengths of our standing desires (both self-interested and altruistic), but also by which desires are emotionally "activated", or occurrently felt, and to what degree. Due to our limited cognitive capacities, we do not — and could not — constantly feel the force of everything that

interests us or that we care about. Their full force is felt only when *triggered*, perhaps by certain thoughts or salient environmental stimuli.

This fact about the variable efficacy of desires shows that not all failures to help others (even holding fixed the magnitudes of the relevant costs and benefits) will necessarily reveal the same insufficiency of good will, and hence the same degree of blameworthiness. In particular, a failure to help others may be more blameworthy (because revealing a greater deficit of good will) in cases where others' needs are especially *salient*, and hence any altruistic desires in the agent can be expected to function at full efficacy. This principle is illustrated in the case of Singer's pond: We feel that a person who could watch a child drown before his eyes must be *unusually* callous. The child's need is so obvious, and so emotionally gripping (for anyone with a modicum of good will), that to fail to act in this case reveals a truly monstrous lack of concern for others.

By contrast, ordinary people with a modicum of good will towards others regularly fail to act in ways that would save the lives of distant strangers. There are many possible explanations of this, but plausibly at least part of the story is that the needs of distant strangers are much less *salient* to us in our everyday lives.[3] A child drowning before our eyes shocks us out of complacency, activating whatever altruistic concern we may have, whereas the constant suffering of the global poor is easier to ignore. (Though, as aid fundraisers quickly learned, a photo and brief life story of a starving child can help to some degree.) This means that our failure to aid the distant does not necessarily reveal a monstrous lack of concern for others. It may be that our moderate concern for others is simply not being activated, and hence fails to guide our actions as it otherwise might.

[3] "Slippery slope" worries may be another common source of inaction, as we worry that helping today would commit us to helping again tomorrow, without any end in sight. In this way, the "expected cost" of saving one may be much higher in the donation case than we might otherwise expect from looking at the case in isolation. Put another way: Saving a unique drowning child is less likely to cause pangs of guilt when we return to normal the next day.

This analysis secures the common-sense result that the Character Evaluation is mistaken: Failure to help the distant needy is typically not as blameworthy as inaction in the case of Singer's pond would be. Moreover, this result is secured on the basis of what, intuitively, seems like the right explanation, namely that *it would take a much worse person* to let a child drown before their eyes, whereas any ordinary non-saint does less than they could (and even, arguably, *should*) for the distant poor.

I think this analysis is interesting for two main reasons. Firstly, it helps to support Singer's argument (and consequentialism more broadly) by showing that we can accept the Act Evaluation without committing ourselves to the implausible Character Evaluation. Second, it shows how facts about what's psychologically *salient* to an agent can alter our assessment of their moral character. In the next section, I'll go on to explore how virtue might sometimes *require* us to find some things more salient than others.

4.2 Virtues of Salience

In recent years, some philosophers have defended the surprising thesis that some virtues essentially involve ignorance or epistemic bias. I will discuss two cases in particular: whether the virtue of modesty involves ignorance, and whether friendship demands that we believe better of our friends than the evidence warrants. In both cases, I will argue, these alleged "virtues of ignorance" are better understood as "virtues of salience" — placing demands on our *attention* and initial *inclinations* to believe, not on our *settled beliefs*.

4.2.1 Modesty and Ignorance

Driver (1989, 1999) argues that the virtue of modesty consists in a disposition to moderately *underestimate* one's own worth. This explains the "Moore-paradoxical" infelicity of the assertion:

1. I am modest.

On Driver's account, the infelicity of (1) is to be explained in terms of the general infelicity of claiming that one's own belief is false: If you think it's false, then how can you believe it? Likewise: If you think you're underestimating your self-worth, then isn't that just to say that you actually think your self-worth is rather higher than previously intimated?

Driver notes that "behavioural" accounts of modesty, either in terms of under-*stating* one's true worth, or a general reluctance to brag, can also account for the oddity of asserting (1). But they fail for the reason that they cannot distinguish sincere from false modesty. The difference, for Driver, is that the genuinely modest person does not merely behave as though she has less worth, she really believes it.

We may wonder: If modesty really involves ignorance in this way, then how is it a virtue? Driver (1989, 383) suggests that it is because modesty-as-ignorance typically arises from "a reluctance to dwell on one's good qualities" or give much thought to rankings, and it is *this* disposition, rather than the resulting ignorance as such, that is truly valuable (at least instrumentally, and perhaps intrinsically as well). But in that case, why not take this valuable disposition, rather than the contingently resulting ignorance, to be constitutive of the virtue of modesty?

Driver (1999, 829) backtracks after considering a case where the two come apart: Albert the scientist puts a great deal of thought into ranking, and then publicly declares himself to be (as he now believes) the fifth best physicist in the world, though in fact the evidence shows that he is third best. Driver bites the bullet and insists that, since he underestimates himself, Albert is modest — albeit in an "anomalous" fashion, "modest *in spite of* his overzealous ranking behaviour." This strikes me as a mistake. It seems much more plausible to think that Albert is not really modest at all, in any normatively significant sense. He is instead (mildly) epistemically irrational, in addition to being immodest. He cares too much about his own rank, and despite

all his efforts he isn't even able to accurately assess what it is. He is, in this way, doubly flawed.

Driver goes on to defend the claim that it is really the ignorance (rather than the anti-ranking disposition) that we value, by considering a case where the two come apart in the opposite direction: Bob knows, on the basis of reliable testimony, that he is the best, though he hasn't himself engaged in any ranking exercise to confirm this. Driver objects that "[a]ny professions of inferiority on his account would constitute false modesty", and hence be found objectionably patronizing and condescending by knowledgable observers. Now, I agree that any such dishonest attempts to placate the presumed jealousy of his audience would constitute condescending rather than modest behaviour. But the mere *possibility* of behaving condescendingly cannot be sufficient to show that an agent is *actually* immodest. (Note that Driver believes Albert, above, to be modest. But he too could behave condescendingly, e.g., by dishonestly reassuring his colleagues that he's "not even in the top ten".)

Suppose that Bob never gives a moment's thought to his relative ranking. His attention is instead directed outward, to opportunities out there in the world, and insofar as he assesses himself at all he does so in non-comparative terms, noticing where he has room for improvement (Brennan 2007). He doesn't think of himself as better than other people, for he doesn't think in terms of comparative rankings at all. (In this way, he differs from the falsely modest person who merely *pretends* not to think of himself as better than others.) In this case, Bob strikes me as a paradigmatically modest person. This is so even though he could, if asked, retrieve from his dusty memory banks the information that the top-ranked person in the world happens to be... him.

On this account — call it modesty-as-salience, in contrast to Driver's account of modesty-as-ignorance — the virtue of modesty need not involve any epistemic error or impairment. It merely requires that the agent not dwell overmuch on her own

ranking or status. She may know the truth of the matter, but it isn't something she *cares* about.[4] And so it isn't something that tends to grip her attention, or intrude into her thoughts.

Both Driver's and my accounts agree that a modest agent will typically not be aware of her ranking or comparative worth. But we diverge when it comes to explaining *why* this is so. Driver proposes that the modest agent must not *believe* the truth about her self-worth, she must instead underestimate it. I instead propose that, rather than giving the wrong answer (like Albert), the modest agent simply does not *attend* to the question.

We can distinguish these two views by employing a "Fate of the World" test. Suppose an evil demon will destroy the world unless Charlie offers an accurate assessment of his abilities. Would Charlie's answering correctly prove that he lacks modesty? On Driver's view, it would. If modesty requires underestimation, then Modest Charlie's best effort at answering the question will yield an incorrect answer. But on my view, this need not be so. Modest Charlie does not dwell on his accomplishments, so the answer may not immediately spring to mind. He may even be *initially* inclined, just as a matter of first appearances, to underestimate himself, due to giving such little thought to his many achievements. But when the stakes are high, he is able to override his characteristic disposition to refrain from self-assessment or ranking behaviour, and assess the evidence in an accurate and dispassionate light. So, upon considering the matter in depth, he is able to give the demon the correct answer. Then, the immediate need having been met, his attention will once again drift away from himself, and back to what he considers to be intrinsically more important matters.

[4] Schueler (1997) offers the related proposal that modesty consists in not caring *whether others are impressed* by you. But as Driver objects, one might disdain others' opinions out of extreme arrogance rather than modesty (though cf. Schueler's (1999) response). My proposal avoids this worry, since the arrogant person is still interested in ranking, it's just that he assumes that others are lower-ranked than he, and disregards their opinions on that basis.

I think that my account of modesty-as-salience has two major advantages over Driver's modesty-as-ignorance view. First, it yields verdicts that are intuitively more plausible in the cases of Albert, Bob, and Charlie, discussed above. Second, I think it is more appealing on theoretical grounds. The only distinctive value of Driver's modesty-as-ignorance would be if we cared a lot that people be able to sincerely reassure us that they aren't so amazing as the evidence indicates. Knowledgeable agents like Bob are unable to do this — any such professions of inferiority from them would, as Driver points out, be insincere and condescending.

But why should we want such reassurances in the first place? Driver suggests that they may be instrumentally useful in defusing problematic social emotions like jealousy that can arise when faced with superior others. But for those others to erroneously claim a lower rank is just one possible way of dealing with this problem. Another, more appealing, solution is for them to refrain from making unnecessary comparative judgments altogether. This is another way that high-achieving agents can defuse jealousy and be less socially threatening. Insofar as their anti-ranking perspective is picked up and shared by those around them, destructive feelings of inferiority may be avoided. So modesty-as-salience shares the instrumental value of modesty-as-ignorance. More importantly, insofar as it consists in the internalization of important moral truths — such as that one's own achievements aren't all that important in the grand scheme of things, that each of us is but one person amongst moral equals, and that comparative rankings lack intrinsic importance — modesty-as-salience has a kind of *intrinsic* appropriateness that befits its status as a genuine *virtue* (rather than merely being a contingently desirable disposition, of no intrinsic moral import).

4.2.2 Friendship and Epistemic Partiality

Keller (2004, 329) uses the following incident from the sitcom *Friends* to suggest that friendship requires epistemic partiality, or thinking better of our friends than the evidence warrants:

Joey and Chandler are playing a game where the latter gives immediate, unreflective and unfiltered answers to the questions asked by the former. Joey has just landed an acting job in Las Vegas, which he hopes will be his big break. He asks Chandler, 'Is this job going to be my big break?', to which Chandler reflexively answers 'No,' putting their friendship in crisis. Chandler's lack of belief in him causes Joey to feel betrayed, and Chandler to feel guilty, suggesting that their friendship involves normative expectations to think well of each other — even in the absence of evidence warranting such optimism.

Stroud (2006, 508) similarly argues:

> [T]he bias of the good friend will normally take the form of casting what she sees or hears in a different light, shading it differently, placing it in a different optic, embedding it in a different overall portrait of her friend. Where our friends are concerned, in short, we become spin doctors.

Both Keller and Stroud conclude that friendship places demands on our *beliefs*. As in the previous section, I want to resist any such strong conclusion, and replace it instead with a more subtle demand on what we find *salient*.[5] I agree with Stroud's characterization of how friendship requires us to *see* things in a more positive light. But I don't see any reason to see friendship as further requiring that we refrain from *correcting for this biased impression* before coming to a settled belief.

Again, we can distinguish the alternative views on offer by means of a "Fate of the World" test: Suppose that an evil demon will destroy the world unless Chandler

[5] Here (and throughout this section) I'm heavily indebted to Helen Yetter-Chappell.

answers correctly whether Joey's new job will be his big break. And suppose that Chandler's initial inclination was (contrary to what was in fact portrayed in the episode) very favourable towards Joey: He's more aware of Joey's strengths than his weaknesses as an actor, which inclines him towards over-rating his friend. And further, his deep hopes for Joey's career link in with the natural human inclination towards wishful thinking, further contributing to Chandler's initial impression that Joey will do well. But then, as the stakes are so high, he pauses and reflects more carefully on the question. He recognizes that, as a friend, he's naturally going to be a bit biased in Joey's favour, so he explicitly adjusts his credences down to correct for this. Further, he tries to take an "outside view", noting that the inductive evidence from Joey's past career failures — not to mention the low base rate of success in the acting business — count against him. Weighing it all up carefully, Chandler concludes (against his initial inclination) that the answer is in fact 'No'.

When his process of belief-formation is spelled out in this way, does Chandler still seem like he has been in any way a bad friend? Surely not. The crucial difference between the original case and this one is that while Chandler still comes to a negative conclusion about Joey's career prospects, in my case his initial inclination was more positive. And this seems to me all that friendship, intuitively, demands. A good friend finds his friend's strengths to be more *salient* than his weaknesses, which naturally leads to an initial inclination towards overestimation. But there is no requirement that we settle for first appearances. We can (if pressed) correct for our biases, and so reach a more accurate final conclusion, without in any way violating the norms of friendship. The problem with Chandler in the original case is not that he *believed* poorly of Joey, but that it didn't even *appear* to him that Joey would do well.

Conclusion

This chapter has explored the two-way relation between salience and fitting character. First, we saw that our evaluations of an agent's moral character need to take into account what they find salient, since neglecting a salient need reflects a greater deficit of beneficent motivation than does neglecting an objectively similar but much less noticeable need. In this way, facts about salience can serve as an important "input" to our moral assessments. But in the second half of the chapter, we saw that the connection also goes the other way: Finding some things more salient than others can be an important "output", requirement, or downstream consequence, of fitting character. In particular, I argued that the virtue of modesty consists in not finding one's own achievements excessively salient, and that a good friend will find his friends' better qualities to be more salient than an impartial stranger would find them. My proposals differ from previously floated views, in both cases, because I insist that there's no further requirement for the agent to *believe* in line with the initial appearances — he may instead correct for any biases introduced by what he finds more or less salient. This compatibility with epistemic norms makes it more plausible that the proposed norms of character are genuine "virtues" or fittingness norms for limited human-like agents.

Bibliography

Adams, Robert Merrihew. 1976. "Motive utilitarianism." *Journal of Philosophy* 73:467–481.

Arpaly, Nomy. 2003. *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press.

Bales, R. E. 1971. "Act-utilitarianism: account of right-making characteristics or decision-making procedures?" *American Philosophical Quarterly* 8:257–65.

Bennett, Jonathan. ms. "Accountability." http://www.earlymoderntexts.com/jfb/accounta.pdf.

Brennan, Jason. 2007. "Modesty Without Illusion." *Philosophy and Phenomenological Research* 75:111–128.

Brentano, Franz Clemens. 1902. *The Origin of the Knowledge of Right and Wrong*. Westminster, A. Constable & co, ltd. Translation of Vom Ursprung sittlicher Erkenntnis. English translation by Cecil Hague.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112:659–688.

Cullity, Garrett Michael. 2006. *The Moral Demands of Affluence*. Clarendon Press.

Darwall, Stephen. 2003. "How Should Ethics Relate to (the Rest of) Philosophy?: Moore's Legacy." *The Southern Journal of Philosophy* 41:1–20.

—. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Dreier, James (ed.). 2006. *Contemporary Debates in Moral Theory*. Malden, MA: Blackwell.

Driver, Julia. 1989. "The Virtues of Ignorance." *The Journal of Philosophy* 86:373–384.

—. 1999. "Modesty and Ignorance." *Ethics* 109:827–834.

Gauthier, David. 1997. "Rationality and the Rational Aim." In Jonathan Dancy (ed.), *Reading Parfit*. Oxford: Blackwell.

Hare, Caspar. 2010. "Take the Sugar." *Analysis* 70:237–247.

Hooker, Brad. 2000. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Oxford University Press.

Hooker, Brad, Mason, Elinor, and Miller, Dale (eds.). 2000. *Morality, Rules, and Consequences*. Edinburgh: Edinburgh University Press.

Hurka, Thomas. 1996. "Monism, Pluralism, and Rational Regret." *Ethics* 106:555–575.

Jackson, Frank. 1991. "Decision-theoretic consequentialism and the nearest and dearest objection." *Ethics* 101:461–482.

Jackson, Frank and Pargetter, Robert. 1986. "Oughts, options, and actualism." *Philosophical Review* 95:233–255.

Kagan, Shelly. 1991. *The Limits of Morality*. Oxford: Oxford University Press.

—. 2000. "Evaluative Focal Points." In Hooker et al. (2000).

Keller, Simon. 2004. "Friendship and Belief." *Philosophical Papers* 33:329–351.

Kelly, Thomas. 2010. "Peer Disagreement and Higher Order Evidence." In Richard Feldman and Ted Warfield (eds.), *Disagreement*. Oxford: Oxford University Press.

Lawlor, Rob. 2009. *Shades of Goodness: Gradability, Demandingness and the Structure of Moral Theories*. Palgrave Macmillan.

Lenman, James. 2004. "Utilitarianism and Obviousness." *Utilitas* 16:322–325.

Louise, Jennie. 2004. "Relativity of value and the consequentialist umbrella." *Philosophical Quarterly* 54:518–536.

Mackie, J. L. 1985. "Rights, Utility, and Universalization." In R. G. Frey (ed.), *Utility and Rights*. Oxford: Basil Blackwell.

Mason, Elinor. 1999. "Do Consequentialists Have One Thought Too Many?" *Ethical Theory and Moral Practice* 2:243–261.

Moore, G.E. 1903a. *Principia Ethica*. Cambridge: Cambridge University Press.

—. 1903b. "Review: The Origin of the Knowledge of Right and Wrong." *International Journal of Ethics* 14:115–123.

Norcross, Alastair. 2006. "Reasons Without Demands: Rethinking Rightness." In Dreier (2006).

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Ord, Toby. 2009. *Beyond Action: Applying consequentialism to decision making and motivation*. Ph.D. thesis, University of Oxford.

—. ms. "How to be a consequentialist about everything."

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

—. 2011. *On What Matters*, volume 1. Oxford: Oxford University Press.

Pettit, Philip. 1994. "Consequentialism and moral psychology." *International Journal of Philosophical Studies* 2:1 – 17.

Pettit, Philip and Brennan, Geoffrey. 1986. "Restrictive consequentialism." *Australasian Journal of Philosophy* 64:438 – 455.

Pettit, Philip and Smith, Michael. 2000. "Global Consequentialism." In Hooker et al. (2000).

Portmore, Douglas W. 2007. "Consequentializing moral theories." *Pacific Philosophical Quarterly* 88:39–73.

—. 2011. *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford University Press.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni. 2004. "The strike of the demon: On fitting pro-attitudes and value." *Ethics* 114:391–423.

Railton, Peter. 1984. "Alienation, consequentialism, and the demands of morality." *Philosophy and Public Affairs* 13:134–171.

Rawls, John. 1999. *A Theory of Justice*. Belknap Press of Harvard University Press, revised edition.

Regan, Tom. 2004. *The Case for Animal Rights*. Berkeley: University of California Press.

Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Belknap Press of Harvard University Press.

—. 2009. "Being Realistic about Reasons." John Locke Lectures.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

Schueler, G. F. 1997. "Why Modesty is a Virtue." *Ethics* 107:467–485.

—. 1999. "Why IS Modesty a Virtue?" *Ethics* 109:835–841.

Shaw, William. 2006. "The Consequentialist Perspective." In Dreier (2006).

Shope, Robert K. 1978. "The Conditional Fallacy in Contemporary Philosophy." *The Journal of Philosophy* 75:pp. 397–413.

Sidgwick, Henry. 1907. *The Methods of Ethics*. Thoemmes Press.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1:229–243.

—. 1993. *Practical Ethics*. New York: Cambridge University Press, second edition.

Slote, Michael and Pettit, Philip. 1984. "Satisficing Consequentialism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 58:139–176.

Smith, Michael. 1994. *The Moral Problem*. Blackwell.

—. 2010. "Beyond the Error Theory." In Richard Joyce and Simon Kirchin (eds.), *A World Without Values: Essays on John Mackie's Moral Error Theory*, 119–139. New York: Springer.

Stocker, Michael. 1989. *Plural and Conflicting Values*. Oxford: Oxford University Press.

Strawson, Peter. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48:1–25.

Stroud, Sarah. 2006. "Epistemic Partiality in Friendship." *Ethics* 116:498–524.

Unger, Peter K. 1996. *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press.

Williams, Bernard. 1973. "A Critique of Utilitarianism." In J.J.C. Smart and Bernard Williams (eds.), *Utilitarianism: For and Against*. Cambridge University Press.

—. 1982. "Persons, Character and Morality." In *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge University Press.

Wilson, Scott. 2006. "Respect for Utilitarianism: A Response to Regan's 'Receptacles of Value' Objection." *Proceedings of the Ohio Philosophical Association* 3.