Utilitarianism and the integrity of the practical realm

Nikhil Venkatesh

UCL

Submitted for the degree MPhilStud Philosophical Studies

Declaration

'I, Nikhil Venkatesh, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.'

Abstract

I investigate Bernard Williams's argument that has come to be knows as 'the Integrity Objection'. Williams gives two cases in which agents are asked to perform some action that is at odds with their commitments, where if they do not perform the action, someone else will, with worse consequences. Utilitarianism recommends performing such actions. Williams's objection is not to this conclusion, but to how utilitarians arrive at it. Utilitarianism regards our commitments as merely one more input into moral deliberation, to be evaluated impartially, and flouted or dispensed with when the utility calculus demands it. Williams believes that we cannot regard our commitments like that; therefore, we cannot deliberate in a utilitarian manner and have commitments.

Adding the premise that utilitarianism recommends that we have commitments, since they make us and others around us happy, Williams makes a charge of incoherence against utilitarianism: it asks us to have commitments, but also to deliberate in a way that makes commitment impossible. One response to this objection is to embrace 'self-effacingness', denying that utilitarianism asks us to deliberate in a utilitarian way. Williams charges that this amounts to utilitarianism 'ushering itself from the scene'. I develop this charge, describing three problems for such a response.

I outline a utilitarian account inspired by Hare, in which there is a time for utilitarian deliberation, and a time for acting from commitments and other non-utilitarian motivations. This account is not wholly self-effacing, and therefore does involve sometimes regarding our commitments in the utilitarian manner Williams thought impossible. I argue that the commitments that are conducive to well-being and therefore recommended by utilitarianism can be regarded in this way without undermining their contribution to well-being. If this account works, therefore, it avoids Williams's objection.

Impact Statement

This thesis aims to contribute to the philosophical discussion of Bernard Williams's 'Integrity Objection' to utilitarianism. It should be beneficial to future research in moral philosophy by providing a schematic reconstruction of Williams's argument, which clarifies its nature and what must be done to resist it; by raising new problems for self-effacing moral theories, and by developing a form of utilitarianism that aims to resist both these problems and Williams's objection.

This thesis addresses the questions of how we should live and how we should think about our choices. Such questions are important to everyone, inside and outside academic moral philosophy.

Acknowledgements

Many thanks to my supervisor, Véronique Munoz-Dardé, and my examiners Alex Voorhoeve and Ulrike Heuer. Mark Kalderon provided useful comments on a draft of section 5.4. Members of the UCL Work in Progress group gave a helpful discussion of Chapter 4. Thanks to Katherine Venkatesh for proof-reading and Hannah Lovell for her patience.

Table of contents

1. Introduction

1.1. Utilitarianism

In our lives, we are faced with choices between different options. An option, as Pettit defines it, 'is a possibility which the agent is in a position to realise or not' (1993, 232), such as one's performing an action. Making such choices is unavoidable. When I wake up, I have the option of showering before breakfast, and the option of eating breakfast before I shower. I must choose one if I am not to remain in bed forever. It is sometimes appropriate to consider such choices from the moral point of view – perhaps I had promised my flatmate that she could shower first. Utilitarianism, as I define it, is the view that as far as morality goes, one should choose the option which will result in the most well-being (that is, the option that will maximise well-being, or is most conducive to well-being).

Utilitarianism enters the philosophical canon with Jeremy Bentham, and was inherited and developed by John Stuart Mill and Henry Sidgwick. (Godwin, Helvétius and Hume are sometimes considered utilitarian forebears of Bentham; Mill even recruits Epicurus to the cause.[1]) Bentham's utilitarianism was radical, taking iconoclastic aim at legal and moral norms inspired by religion and tradition. By the end of Mill's life, utilitarianism had a widespread influence on government policy, especially in Britain and its empire. As Wiggins notes:

> 'it involved itself… in campaigns for law reform, prison reform, adult suffrage, free trade, trade union legislation, public education, a free press, secret ballot, a civil service competitively recruited by public examination, the modernisation of local government, the registration of titles to property in land, safety codes for merchant shipping, sanitation, preventive public medicine, smoke prevention, an Alkali Inspectorate, the collection of economic statistics, anti-monopoly legislation… In sum, philosophical utilitarianism played

[1] Mill's idiosyncratically inclusive canon of utilitarian thinkers also includes John Brown and Samuel Johnson, and unnamed thinkers 'long before' Epicurus. Indeed, he writes: 'In all ages of philosophy one of its schools has been utilitarian'. (1971, 54)

> a leading part in promoting indefinitely many of the things that we now take for granted in the modern world.' (2006, 145)

Throughout the twentieth century a version of utilitarianism underpinned mainstream economics, and through it influenced policy across the world. Philosophers, though, became ever more suspicious of it. And as a personal rather than a public philosophy (Goodin 1995) – an answer to the question of what individuals rather than public policy ought to do – utilitarianism has never been widely accepted. It is in this personal role that this thesis considers utilitarianism.

Utilitarianism is a species of consequentialism, which holds that one should choose the option which will result in the best consequences. The distinctive feature of utilitarianism is that it holds that the best consequences are those instantiating the greatest sum of well-being. Its appeal stems from the widespread conviction that well-being matters in the way that whatever underpins morality must matter. As Scanlon puts it: 'It seems evident to people that there is such a thing as individuals' being made better or worse off. Such facts have an obvious motivational force; it is quite understandable that people should be moved by them in much the way that they are supposed to be moved by moral considerations.' (1982, 108) Of course, most people think that other things matter as well. But well-being seems a particularly attractive basis for morality, for several reasons. Firstly, a belief in the importance of well-being is *ubiquitous*. Some people think football matters, and some think it doesn't; some think music matters, and some are tone deaf; some think honouring the gods matters, others are atheists: everyone agrees that well-being matters. If we hope to develop a morality that can be justified to everyone, well-being therefore seems a good place to begin. Secondly, well-being is indisputably *real*. Moralities that give primacy to rights, principles, virtues and gods are vulnerable to the charge that they fetishise the contrivances of philosophers or of ideology. Whereas well-being can be felt and perceived (especially if it is identified with pleasure minus pain, as in Bentham and Mill) the elements of morality mentioned above seem to be 'mere ideas, without any foundation in reality.' (Scanlon 1998, 152) How could they imbue morality with the significance we take it to have? Thirdly, many other things that matter to people seem to be *explicable* in terms of well-being. Football and music matter to those who enjoy them; religious devotion to people who become, or think they will

become, happier from it. And rights, virtues and so on – at least the ones that most people agree are good – seem to be such that it is conducive to well-being that they are observed. If the significance of well-being can explain why all these things matter to us, this suggests that well-being determines moral properties, just as if the all physical properties can be explained by the arrangement of atoms we might infer that atoms determine physical properties. Once the primacy of well-being is established, the appeal of utilitarianism is the appeal of taking a consequentialist line towards it. This can seem irresistible. If well-being is good, then the moral thing to do is to promote it. In prudential reasoning, we would rarely choose something that made us worse, rather than better off. In moral reasoning, we should therefore choose the options that make things in general better – to do otherwise would violate the almost tautologous-sounding maxim that we should 'make the world as good a place as possible' (Scheffler 1988, 1).

These remarks fall far short of proving utilitarianism, or even offering a compelling argument for it. They merely show that utilitarianism has some prima facie appeal. This appeal, together with its canonical status and historical importance, makes utilitarianism – although a minority position in contemporary philosophy – 'a position one must struggle against if one wishes to avoid it.' (Scanlon 1982, 103) One method of struggle has been to demonstrate its variance with 'common sense' moral beliefs and practices: for example, by showing that it would require us to push fat men in front of trains, give no favour to our children, and harvest the organs of healthy, living patients. Another is to assert that there are values other than well-being and not explicable by it: equality, love, faith and so on. Another is to develop and advocate a rival moral theory, as Scanlon has done (1982, 1998).

In what is known as 'the Integrity Objection', Bernard Williams (1973) struggles against utilitarianism in a different way. As I interpret him, he aims to show that utilitarianism contradicts not only commonsense moral beliefs and practices, but itself: utilitarianism is, in his view, incoherent. In this thesis, I will reconstruct Williams's argument and one response utilitarians make to it. I consider some problems with this response, and end by sketching a utilitarian account that navigates a way between Williams's argument and those problems.

1.2. Williams's cases

Williams's objection is made through two hypothetical cases (1973, 97–99). In one, a recently graduated chemist, George, is offered a job in a chemical warfare laboratory. He decides that he cannot accept, since he is opposed to chemical warfare. He cannot accept even though his unemployment causes him and his family to suffer, and even when he is told that the person who would be hired in his place would pursue the research in such a way that more dangerous chemical weaponry would result. In the second case, Jim, in a foreign land in the aftermath of an uprising, is made an offer by Pedro, an army captain. Pedro will execute twenty innocent prisoners as a warning to dissenters unless Jim agrees to shoot one himself, in which case the other nineteen will be released.

The fact that Williams uses hypothetical cases invites the thought that he wages his struggle against utilitarianism in the first way mentioned above: showing that its recommendations are at odds with common sense moral beliefs and practices.[2] Utilitarianism recommends, given some additional assumptions (that George wouldn't be so depressed by taking the job that he and his family are caused more suffering; that the development of chemical weapons is bad for well-being; that Pedro's prisoners have lives worth living) that George takes the job and Jim shoots the prisoner. Those who believe that one should never assist with research into chemical warfare, or kill, or 'sell out one's principles' will disagree. These recommendations alone will persuade such people against utilitarianism. But Williams does not want to persuade only people with such moral beliefs – and is not one of them himself, remarking that 'the utilitarian is probably right' in Jim's case (1973, 117). The cases are meant to make salient a certain feature of moral life, consideration of which reveals utilitarianism to be defective. That feature is integrity, and the defect Williams points to is incoherence.

One may ask whether utilitarianism must recommend that George and Jim accept their offers. What is typically called 'rule-utilitarianism' holds that we should act in accordance with the set of rules whose general adoption would maximise well-being. It seems plausible that 'do not assist with research into chemical warfare' and 'do not kill innocent prisoners' would

[2] This is Hare's interpretation of Williams (1981, 49, 130–46).

be part of this set. In that case, rule-utilitarianism would recommend that George refuse the job, and Jim refuse to shoot.

Rule-utilitarianism, however, is not a kind of utilitarianism as I have defined it. Above, I said that utilitarianism is the view that as far as morality goes, one should choose the option which maximises well-being. Rule-utilitarianism applies this principle to the choice of rules, but then applies a different principle – choose the action that conforms to the rule – for actions. As I define utilitarianism, this principle is applied to all options, that is, every possibility that an agent is in a position to realise or not. One kind of option is which actions the agent performs. Another is which rules they adopt. Another would be which dispositions they cultivate and which projects they pursue. So it may be that although utilitarianism recommends adopting the rule 'do not kill innocent prisoners'; it will also recommend that Jim breaks it, if doing so would lead to more well-being.[3]

Wiggins finds the expansiveness of utilitarianism under definitions like mine suspect. He defines utilitarianism more narrowly, as the view that seeks 'to fix the extension of the predication "acts rightly" purely in terms of consequences' (2006, 149): specifically, that one acts rightly when one performs the action whose consequences involve the most well-being. He complains that Mill

> 'proceeds as if, in addition to propounding a principle of utility for the rightness of actions, he is also proposing a principle of utility for the evaluation of all sorts of other items, namely policies, practices, measures, reforms, rules, etc. – a principle that is confined neither to rightness nor to actions.' (2006, 152)

For Wiggins, this leads to an unstable 'double-mindedness' (2006, 153, 163): his reasons for this conclusion are similar to Williams's which will be the focus of this thesis. My definition of utilitarianism is as expansive as that of Wiggins's Mill: it gives a principle for choosing options, whether those options are actions, rules, dispositions, policies or anything else that an

[3] I will consider in Chapter 4 a utilitarian position that claims that in cases like George's and Jim's agents might not truly have the option of accepting their offers, or that accepting might be possible only as part of an option which will not lead to the most well-being – in which case utilitarianism would not recommend acceptance.

agent can realise. In addition, my definition of utilitarianism says nothing about the extension of 'acts rightly'. It focusses on the question of what we should do, and it is not absurd to think that these things might come apart. I will consider this thought at greater length in section 4.4.

With utilitarianism defined and Williams's cases introduced, we can now move onto the substance of the Integrity Objection.

2. The Integrity Objection: incompatibility

2.1. Responsibility and projects

Williams introduces his discussion of integrity by considering

> 'the idea, as we might first and simply put it, that each of us is specially responsible for what *he* does, rather than what other people do. This is an idea closely connected with the value of integrity. It is often suspected that utilitarianism makes integrity as a value more or less unintelligible. I shall try to show that this suspicion is correct.' (1973, 99)

He goes on,

> 'the reason why utilitarianism cannot understand integrity is that it cannot coherently describe the relations between a man's projects and his actions.' (1973, 100)

We can make out Williams's claim of a 'close connection' between integrity and responsibility, and hence his notion of integrity, in terms of these relations. There is, for Williams, a deep difference between my relations to my actions and to those of other people, even when I can prevent or encourage the latter. This difference is at play in the cases of George and Jim:

> 'The situations have in common that if the agent does not do a certain disagreeable thing, someone else will, and in Jim's situation at least the result, the state of affairs after the other man has acted, if he does, will be worse than after Jim has acted, if Jim does. The same, on a smaller scale, is true of George's case.' (1973, 108)

Jim could shoot one person or reject Pedro's offer in which case Pedro will shoot twenty. George could advance the manufacture of chemical weaponry or reject the job, allowing his rival to advance it in more dangerous directions. Whether the second state of affairs is realised, in both cases, is in the gift of Jim and George. It is, in the sense introduced above, an option for them. But it would be misleading, according to Williams, to think of those two men as having brought about these states of affairs if they are realised. It would be misleading to think that Jim and George will merely

have had an 'effect on the world through the medium… of Pedro's [or the unnamed rival chemist's] acts.' (1973, 109) In Williams's view their responsibility for those states of affairs is therefore lesser and qualitatively different; morality respects a distinction between my actions, and actions that are not mine but whose occurrence I have control over. What could account for this distinction? Williams points to the relationship between actions and projects. If George takes the job, he adopts the development of chemical weapons as a project, and accordingly conducts the relevant research. If he doesn't, the other chemist would adopt the same project, and pursue the same research, but there would not be the same relationship between *George's* projects and the research. (Indeed, George would retain his project of opposing chemical warfare.) If Jim were to reject Pedro's offer, twenty people would die. But this would not be because Jim had a project that aimed at their deaths, but because Pedro did. Their deaths in this case would thus be best described as a killing by Pedro, not by Jim, for all the opportunity Jim had to save them. We are 'identified', as Williams says (1973, 116), with the actions that flow from our projects.

What if Jim were to accept Pedro's offer? Although Jim would pull the trigger, it would seem wrong to hold him responsible. Pedro's coercion effectively turns *Jim* into a medium through which *Pedro* affects the world. This intuitive description of the case is reflected at the level of projects: the killing is the aim of Pedro's projects, not Jim's. When we perform actions which do not flow from our projects – as Jim does here – we are not identified with them: rather, we are alienated from them. Our responsibility for such actions is attenuated.

That 'each of us is specially responsible for what *he* does', then, seems to mean this: each of us is specially responsible for the actions that flow from our projects. They must flow from our projects in the right way: if my pursuing some innocent project outrages you so much that you lash out at me, I am not responsible for this, even though my project causally preceded your lashing out.[4] Rather, the actions must flow from our projects in a way that is directed by an aim of those projects, like Jim's shooting of a prisoner would be directed by Pedro's aim of intimidating dissenters. This – not the fact that we had the opportunity to determine whether the action was performed or not – is what makes those actions ours and not someone

[4] Thanks to Véronique Munoz-Dardé for this example.

else's. To neglect this connection is to attack our integrity. 'Integrity' here is meant in the sense of wholeness or unity – an agent's integrity is the unity between them, their projects and their actions.

Utilitarianism seems to neglect this connection. As I have defined it, utilitarianism provides a criterion of choice between options: what one should do, according to utilitarianism, is determined by the effects on well-being of each option available to you. An agent's options are whatever they are able to realise: this is not limited to actions flowing from their projects. Furthermore, utilitarianism is indifferent between different paths to the same sum of well-being, as reflection on Williams's case of Jim demonstrates.

Jim has two available options: (1) accept the offer and shoot one prisoner; (2) reject the offer and see Pedro shoot twenty. Choosing (1) will lead to more well-being than choosing (2), in normal circumstances (the prisoners will have lives worth living if they survive, the bereaved will have reduced well-being, and so on). So utilitarianism recommends (1). But notice that the very same reasoning would apply if Pedro were not in the picture. Imagine Jim had a choice between (1) shooting one person and (2') shooting twenty people himself. The effects on well-being are equivalent (except perhaps for differences in guilt felt by Jim and Pedro) across (2) and (2'). For utilitarianism, therefore, these options are equivalent, and Jim's choosing (2) is as bad as his choosing (2') would be. This is so even though in (2) the killings would flow from Pedro's projects, and in (2') they would flow from Jim's. Utilitarianism therefore seems to neglect the moral significance of the identification of an agent with their actions through their projects.

What is a project for Williams? He gives no explicit definition. He gives examples (1973, 110–11): desires for oneself, one's family and one's friends to have the basic necessities of life, and for the 'objects of taste'; 'pursuits and interests of an intellectual, cultural or creative character'; political causes such as Zionism; 'projects that flow from some more general disposition towards human conduct and character, such as a hatred of injustice, or of cruelty, or of killing'; the utilitarian project of maximising well-being. A project, to be something from which action may 'flow' (Williams's preferred verb for this relation), must be capable of motivating the agent who has it. The motivational aspect of projects is reaffirmed by

Williams when he says that if utilitarianism demands that we 'step aside' from our projects, we are alienated 'from [our] actions and the source of [our] actions in our own convictions.' (1973, 116) But projects cannot be *whatever* motivates action – a habit or addiction would not be a project. We are conscious that our projects guide our actions (unlike mere habit) and approve of them (unlike addictions).[5,6]

Desires for basic necessities motivate us all, but political causes or moral convictions are more individual: one might or might not be motivated by Zionism or justice. These things are projects for some people, but not for others. But all projects must be such that they are had in some unique way by each individual who has them, to do the work Williams puts them to in explaining integrity and responsibility. If there are two people who have Zionism as a project, and one, motivated by that Zionism, performs some action, then the other is not responsible for that action as if it were theirs. (We might think that they have some responsibility to condemn such an action if it is wrong or to defend it if it is right; perhaps one cannot think that actions done by others for the sake of a project you share are none of your business. But Jim also, as Williams says (1973, 110), cannot take Pedro's actions to be none of his business, and this does not make those actions his.) This is needed to distinguish George's taking the job – and hence pursuing an evil project – from the other chemist's taking the same job and pursuing the same project. So, when he says that each of us is specially responsible for the actions that flow from our projects, Williams means that each of us is specially responsible for the actions that flow from *our having* our projects, not for actions that flow from other people's having projects we happen to share.

---

[5] Projects, as I think of them, are similar to Joseph Raz's 'goals'. (1986, chap. 12)

[6] If desires are projects, does that make Jim's action unalienated? If Jim were to accept Pedro's offer and shoot, it might be argued, he would in some sense have a desire to pull the trigger and kill his unfortunate target. He would be conscious of this desire and approve of it as part of his least bad option. However, it seems clear to me in such cases that the most important motivation driving Jim's action would be Pedro's strategic and sadistic projects. So we might say that anything that motivates action can be a project, and our actions are alienated if it is not our projects that are the primary motivation for them.

2.2. Utilitarianism and commitments

These two passages form the crux of the Integrity Objection:

> 'how can a man, as a utilitarian agent, come to regard as one satisfaction among others, and a dispensable one, a project or attitude round which he has built his life, just because someone else's projects have so structured the causal scene that that is how the utilitarian sum comes out?'
>
> 'It is absurd to demand of such a man, when the sums come in from the utility network which the projects of others have in part determined, that he should just step aside from his own project and decision and acknowledge the decision which utilitarian calculation requires.' (both passages Williams 1973, 116)

In the second quoted passage Williams contrasts the agent's 'own project and decision' with the utilitarian recommendation. This is a false dichotomy. An agent could adopt utilitarianism itself as a project – indeed, Williams himself considers this possibility just one page before. For this agent, acting as the utilitarian calculation requires *would* count as acting from their own project and decision. The action would flow from one of the agent's projects, so this would not be a case of alienation, in which the action was not really the agent's own. As long as such an agent is possible, there is no necessary opposition between unalienated action from one's own project and decision and acknowledging utilitarian recommendations.

However, as the cases of George and Jim show, utilitarian recommendations can conflict with other, non-utilitarian projects. They cannot simultaneously follow the utilitarian recommendation and their projects of opposing chemical warfare and refraining from killing. This may be a problem for George and Jim but is not obviously a problem for utilitarianism. We often have multiple projects such that on some occasions not all can be followed. I might want to go to a friend's birthday party, as part of an ongoing project of friendship, and at the same time to play cricket, as part of an ongoing project of improving my sporting ability and fitness. The diary clash means that I cannot do both but does not mean that one of those projects is defective in some way. Williams's objection cannot be simply that following utilitarianism leads to such dilemmas with other of our projects.

His objection, I think, is deeper: insofar as the agent acts on utilitarianism, their attitudes towards their other projects are defective. This is because of the impartiality with which utilitarianism regards projects. Say that I have some project, and a stranger has a different project. Both will increase the sum of well-being in the world, but to different extents. How are they to be valued? According to utilitarianism, projects are important insofar as they are conducive to well-being, impartially conceived. For utilitarianism, one project is no more important than another just because of the person who has it: in deciding what to do, my project carries some weight, but so does the stranger's; the value of both depends on the same basis, and the fact that it is mine is no reason to think that mine has greater weight. I should act on the project which is such that doing so leads to more well-being, whether that is mine or the stranger's. But if I deliberate in this way, how is my project *my* project? It seems obvious that if X is my project and Y is not, I must regard X in a different light to Y (typically as more valuable) and be generally disposed to act on X rather than Y. So for the agent with a utilitarian project and some others, the latter seem to have a double life: they are both that agent's projects, special to her, and they are, according to utilitarianism 'one satisfaction among others'. The utilitarian agent's actions do not flow from these projects, but rather from well-being calculations that take into account everyone's projects equally. The utilitarian agent is therefore somewhat (though not wholly, if utilitarianism itself is one of their projects) alienated from her actions. This, I think, is what Williams means by alleging that utilitarianism 'cannot coherently describe the relations between a man's projects and his actions'.

Secondly, in this passage, Williams alleges an absurdity in demanding that someone step aside from their projects. If to 'step aside from' a project is simply to perform some action antithetical to it, then what seems absurd to demand is that morality never ask one to step aside from one's projects. A project, as we have seen, could be a simple desire or taste. There are surely occasions in which we ought to forego satisfying one of our desires to help someone else satisfy theirs (consider Singer's (1972) case in which a child's life can be saved at the cost of some muddy clothes). This is an essential part of morality. I think Williams agrees with this. He writes that 'in the case of many sorts of projects' it is 'perfectly reasonable' to weigh the utility gains of your satisfying your project against the gains of someone else satisfying theirs when the two conflict (1973, 115–16). This not only permits a moral

theory to ask us to abandon our projects on occasion, it also affirms the utilitarian method of counting one's projects as 'one satisfaction among others'. If the previous paragraph is correct, Williams believed that thinking in such a way was incoherent and somewhat alienating, but perhaps reasonably so. His objection to utilitarianism, therefore, is not that it sometimes asks us to act at odds with our projects for the sake of utility – for some sorts of projects, this is as it should be.

Not, however, according to the first quoted passage, for projects around which one builds one's life. Williams is especially interested in this subset of projects, which he calls 'commitments'. What distinguishes commitments from other projects is left vague, but has to do with the greater strength of the attitude one has towards them, hinted at by words like 'thorough', 'deep and extensive' and 'serious'. 'One can be committed', Williams writes, 'to such things as a person, a cause, an institution, a career, one's own genius, or the pursuit of danger.' (1973, 112) A commitment is not simply a very strong desire, though; it is a project which in some way defines the person who has it. Consider the desire to eat: when one is very hungry it may be overwhelmingly strong, but it is hardly something that defines one's character and shapes one's life. Williams writes that one could treat a cultural pursuit as a commitment. One's relationship to that pursuit would be 'at once more thoroughgoing and serious than their pursuit of various objects of taste, while it is more individual and permeated with character than the desire for the basic necessities of life.' (1973, 111) Enjoying the tune of some aria does not count as a commitment, even if it motivates you to go to an opera. Being an opera-lover, on the other hand, which involves educating oneself about the history and subtleties of the form, keeping oneself abreast of current productions, watching and listening to opera frequently, defending its value in argument, and so on, could be a commitment. Insofar as there is a distinction between an opera-lover and someone who enjoys the opera, it seems that for the former their relationship with opera has permeated their character, such as to become partly constitutive of their identity. If being an opera-lover is related to us in this way, and essentially involves certain actions, then performing those actions is essential to our being who we are. This means that a different level of integrity is at stake in the actions flowing from our commitments. Actions flowing from our projects are ours; actions flowing from our commitments are not only ours, they are us.

A person who has these sorts of projects cannot see them as 'one satisfaction among others'. As Williams puts it concerning a subset of commitments, moral convictions: 'we… cannot regard our moral feelings merely as objects of utilitarian value… to come to regard those feelings from a purely utilitarian point of view, that is to say, as happenings outside of one's moral self, is to lose a sense of one's moral identity; to lose, in the most literal way, one's integrity.' (1973, 103–4) For utilitarianism, the fact that Jim has a moral conviction against killing is just another input to the calculus, like the fact that one of the prisoners at risk has a young child whose life will be made substantially worse if he dies. And it will only be a significant input insofar as it entails that breaching this conviction will make Jim unhappy. But Jim cannot view his conviction like that, as one additional input to a sum, which matters only because of its causal effect on his mental states. We value our convictions because we believe they are true, important and part of who we are. Williams writes that 'once he is prepared to look at it like that [as an input to a utilitarian sum], the argument in any serious case is over anyway.' (1973, 116) I take this to mean that the utilitarian way of thinking threatens our capacity to be committed to their moral convictions.

Raz, inspired by Williams, argues that there are some goods and relationships one cannot have without regarding them in a different way to how that utilitarianism regards them. These are what he calls 'constitutive incommensurabilities' (1986, 345–57). A plausible case is friendship. Raz writes: 'Only those who hold the view that friendship is… simply not comparable to money or other commodities are capable of having friends.' (1986, 352) To be a friend is, in part, to find it 'abhorrent' (1986, 346) to be asked to make a trade-off between one's friend and some monetary reward, however large. Utilitarianism countenances such trade-offs: if I were offered a million pounds to never see a friend again, considering the good that money could do, it is plausible that I should accept. Everything is comparable, for utilitarianism, in a common currency: well-being. For any two options, one is more choiceworthy than the other insofar as one is more conducive to well-being and the other less.

If commitments are commensurable with other goods, as utilitarianism asserts, then utilitarianism requires that one countenances abandoning them. For any commitment – no matter how much it means to you – the utilitarian conceives that the causal scene could be structured such that

abandoning it is the thing you should do. Another agent or the natural state of the world could always raise the utility costs of fulfilling and maintaining one's commitments to a level at which dropping them would be more conducive to well-being. So to regard a commitment as 'one satisfaction among others', given that different satisfactions are commensurable, is to regard it as 'dispensable'.

We saw above how the projects of a utilitarian agent lead a double life: they specially motivate the agent and distinguish her actions from those of others, but the agent (as a utilitarian) places no greater moral weight on them than on anyone else's. I take Williams to make the claim that commitments cannot have such a double life: if an agent has a commitment to something, they cannot value it in the utilitarian way. That they 'cannot', rather than 'should not', is important. A 'should not' would simply invite the utilitarian to dispute Williams's moral intuition. They could say, 'Well, I think that agents *should* regard their commitments in such a way.' Williams does not use straightforwardly normative language in this part of his critique. He asks, 'how *can* a man…' and calls the utilitarian demand not wrong, but 'absurd'. This is meant to block such a response. The suggestion is that it is impossible to have a commitment and regard it in a utilitarian manner. Agents cannot view their commitments as 'happenings outside of one's moral self', as comparable to other goods which are, or as dispensable 'when the sums come in'. If I were to regard my relationship with someone as equivalent to some amount of money, it could not be a committed relationship, like a good friendship or marriage. If I were to regard opera as valuable just insofar as it made people happy, rather than for its own sake, I would not be an opera-lover. If I were to regard my Zionism as one more input into finding the optimific resolution to the situation in Palestine, rather than my 'way of seeing the situation' (Williams 1988, 190), bound up with who I am, I would not be a committed Zionist. The agent who looks at things in a utilitarian way cannot have such commitments.

To take stock: the Integrity Objection does not imply that the actions of a utilitarian agent would be wholly alienated, since utilitarianism can itself be a project; nor does it claim that the flaw in utilitarianism is that it sometimes asks us to act contrary to our projects. Williams believes that regarding one's non-utilitarian projects in a utilitarian way involves some incoherence and alienation, but he judges that this can be reasonable. The crux of the

objection is that utilitarianism asks us to regard our *commitments* – the projects that define who we are – in a way (impartially and as dispensable) that is impossible for us insofar as we have commitments at all.

2.3. Robustness and incapacity

So far, the Integrity Objection has been reconstructed in terms of thought: how utilitarianism requires us to regard our commitments, and how commitments, according to Williams, must be regarded. But commitments seem to require action as well as thought (one cannot be an opera-lover without going or listening to the opera, or a friend to someone without ever lifting a finger to help them). Utilitarianism is also likely to conflict with commitments on this score: there are always conceivable situations in which it recommends actions which are incompatible with the actions required by some commitment. I might have a choice between going to see a friend in hospital, which would honour our friendship but not be much fun for either of us (he would not be good company for me, and my presence would make him feel guilty or jealous) and going to watch a cabaret which would be highly enjoyable. Utilitarianism recommends the latter. A committed vegetarian, presented with meat by an easily upset host, might be asked by utilitarianism to eat it and appear to enjoy it. Whilst, as I said above, it is not a blow to utilitarianism that it sometimes asks us to perform actions contrary to our desires, it is more worrying that it may ask us to act at odds with our commitments. The permeation of character by commitments means that actions stemming from commitments are part of how we maintain our distinctive selves – our integrity, in the sense of the term associated with wholeness and unity. However, once again we can point out that utilitarianism itself could be a commitment. There are people, for example in the 'effective altruist' movement, of whom it can be said that they have built their lives around the maximisation of well-being. For them, if utilitarianism asks us to act contrary to another commitment this is a mere conflict between commitments. Not only do such conflicts seem possible without rendering either commitment defective, that we may sometimes breach one commitment for the sake of other moral considerations is entailed by Williams's judgment that Jim should accept Pedro's offer. So his objection cannot be simply that utilitarianism sometimes requires action at odds with other commitments.

If utilitarianism prevented us from having other commitments at all, however, that would be a problem. That is what, according to my reconstruction of his argument in the previous section, he claimed with respect to thought – utilitarianism makes us think about commitments in a way that makes it impossible for us to have them. Does it also make us act in a way that precludes them?

There is a second meaning of integrity, roughly synonymous with sincerity and incorruptibility; someone is 'a person of integrity' insofar as they stick to their principles in a range of circumstances. A judge with integrity, for example, will deliver fair trials, however much money she is offered to do otherwise. This demonstrates her commitment to justice. Pettit identifies 'robustly demanding goods', which require 'the provision of a less demanding benefit not just actually but across a range of possible cases' (2015, 13). Thus, a judge who provides justice must not only provide fair trials, but cannot be someone who *would* accept a bribe – even if they never do because they never receive an attractive enough offer. Shakespeare's Sonnet 116, quoted by Pettit, claims that 'Love is not love, which alters when it alteration finds.' The suggestion is that to love someone is not merely to care for them in certain ways, but for it to be the case that if things were different, one would nevertheless care for them in the same ways. A true lover is prepared to stick with their beloved through thick and thin ('for richer, for poorer; in sickness and in health…'), and if they would not they do not really provide love.

It is tempting to interpret Williams's claims about integrity and utilitarianism in the light of Pettit's observations. The claim would be that having a commitment is robustly demanding across possibilities in which other relevant values are the same, but the well-being sums are different. Thus, if George has a commitment to opposing chemical warfare, and this demands that he refuse to participate in research that will further its efficacy, then he will refuse to accept the job no matter how much money is offered. If a £50 000 salary is offered rather than a £20 000 one, then that will tip the balance of well-being further towards George accepting the job. But because his commitment to opposing chemical warfare is robust, this will not alter his decision. He cannot accept, whatever the sums at stake. That is why it is 'absurd to demand' that people step aside from their commitments 'when the sums come in'.

This notion of commitments as robustly demanding relates to what Williams called 'moral incapacity'. This is 'the kind of incapacity that is in question when we say of someone, usually in commendation of him, that he could not act or was not capable of acting in certain ways.' (1992, 59) George, as Williams puts it, 'cannot' take the job. The idea is not simply that he should not, nor that he will not in these circumstances – although both of these are true – but that it is not possible for him. If it were, we could not say that he had the commitment to oppose chemical warfare. This impossibility is not unbounded. If George signed the contract when a gun was put to his wife's head and a pen in his hand, one could not say that this undermines his claim to be committed to opposing chemical warfare. Williams writes: 'It is plausible to say, with the pessimist, that if having a moral incapacity implies that there are no circumstances at all in which the agent would knowingly do the thing in question, then there are no moral incapacities. Ingenious coercion or brutal extremity can almost always produce such circumstance.' (1992, 69)

What are the bounds of moral incapacity, and hence the demands of robustness? For Williams, a moral incapacity is at least 'proof against rewards' (1992, 69) – if a greater salary is offered, George will still refuse the job. This aligns with Pettit's notion of robustness and the ordinary use of integrity in its second sense. Utilitarianism is not proof against rewards. There is no option that utilitarianism will rule out in all possible worlds that differ only in the rewards on offer, since in some the rewards will be high enough to outweigh, in terms of well-being, any bad that comes of accepting. What this means is not just that utilitarianism will sometimes ask us to act contrary to our commitments, but that utilitarianism will never allow us to act in accordance with our commitments with the robustness necessary for us to count as having them. Therefore, utilitarianism seems incompatible with our having other commitments.

The claim that having a commitment involves robust action goes further than the claim that having a commitment is incompatible with regarding that commitment as dispensable. One can regard a commitment as indispensable, whilst it remains the case that one would drop it in the right circumstances. People make marriage vows with great sincerity, regarding them as sacred, and then break them when a pretty stranger comes along making what seems to them a better offer. Until that stranger arrived on the

scene, they may well *regard* their vows as non-negotiable. But their performance of the actions required by the vows, even before the stranger's arrival, would not have been robust in Pettit's sense. Could such a person be said to have ever been committed to their spouse?

It is not obvious that Williams thought that having a commitment necessitates robustness in action. In *A Critique of Utilitarianism* he usually talks about how agents regard their commitments, and how they relate to their character, rather than how they act. But robustness would explain the absurdity he find in the utilitarian demand that we step aside from our commitments. It is also one way of explicating the 'thoroughness' commitments have compared with desires and tastes. If it would benefit me, or others, I might forego fulfilling some desire or enjoying some object of taste. But there are fewer possibilities in which I would fail to honour a commitment. Lastly, Williams's later work on moral incapacity shows that he did think that such robustness was valuable, whether it is necessitated by commitments or not. I will therefore take this notion of robust action to form part of Williams's Integrity Objection to utilitarianism. If commitments require such robustness, and utilitarianism is not robust against changes in rewards, the utilitarian agent could not have commitments.

2.4. The objection so far

Williams's argument as I have reconstructed it so far aims to show that one cannot follow utilitarianism and have commitments. Utilitarianism does not place enough weight on the distinction between actions flowing from my projects and those flowing from others' projects, and for this reason gives our projects an incoherent and somewhat alienated double life in our deliberation. Such deliberation may be unobjectionable for some projects, but not for commitments. Utilitarianism, the objection goes, requires us to regard our commitments from an impartial standpoint, as fungible in the currency of well-being and therefore dispensable. It also requires us to act contrary to our commitments, if the reward is high enough, preventing us from honouring our commitments robustly. Therefore, if (as Williams claims) we must regard our commitments as part of us and indispensable,

and if they demand actions robustly, having commitments is incompatible with following utilitarianism.

3. The Integrity Objection: incoherence

3.1. From incompatibility to incoherence

Under my interpretation the Integrity Objection is a charge of incoherence. We can put Williams's argument as discussed in the previous chapter thus:

1. Having a project as a commitment
    a. is incompatible with regarding that project impartially or as dispensable;
    b. requires certain action robustly with respect to changes in rewards.
2. Utilitarianism requires us to
    a. regard all projects impartially and as dispensable;
    b. act in a way that is sensitive to rewards.
3. Therefore, utilitarianism requires us not to have commitments.[7]

So far, this is a charge of incompatibility, not incoherence. The gravity of the charge depends on the importance one assigns to commitments. Williams thought that they were very important. But a utilitarian might differ. The objection at this point has a familiar form: some non-utilitarian moral principle is at odds with utilitarianism, and if we value the former we cannot subscribe to the latter. Similar objections to utilitarianism claim an incompatibility with equality, rights, virtues, and so on. Such objections may be compelling, but the defender of utilitarianism can always 'bite the bullet' on them: sticking by utilitarianism and recommending that we give up on equality, rights and virtues.

The same response cannot be given to a charge of incoherence; an incoherent theory is at odds with itself, so there is no other value to give up on. If utilitarianism is incoherent, that is sufficient reason to reject it. This is not because we ought not to believe inconsistencies. An incoherent moral theory could not give us advice or evaluate our actions, since what we should

[7] This conclusion follows given the principle that if a moral theory requires us to do something, and doing that thing is incompatible with doing some other thing, the theory requires us not to do the second thing. Note that for 3 to be true, it is sufficient for either 1a and 2a or 1b and 2b to be true.

do, according to such a theory, is indeterminate. It would therefore fail to fulfil the function of morality.

Williams considers the bullet-biting utilitarian response: 'perhaps, as utilitarians sometimes suggest, we should just forget about integrity, in favour of such things as the general good.' (1973, 99) He thinks that such a response, if he is right, is inadequate, 'since the reason why utilitarianism cannot understand integrity is that it cannot coherently describe the relations between a man's projects and his actions.' (1973, 100) We have already seen one incoherence that Williams identifies: the double life of our projects, as both specially motivating for us and one factor amongst other equally important considerations in a utility calculation. But it cannot be this incoherence that condemns the utilitarian, because Williams admits that regarding projects in this way is sometimes reasonable. The incoherence I think Williams ultimately finds in utilitarianism has to do with commitments. The idea is that commitments are important to well-being, and therefore utilitarianism must value them – whilst it makes it impossible for us to have them. This interpretation accounts for his promise to show that utilitarianism 'can make… only very poor sense of what was supposed to be its own speciality, happiness.' (1973, 82)

Adding a further premiss, the incompatibility argument above can be made into a charge of incoherence. The premiss is that utilitarianism should require us to have commitments (other than to utilitarianism). If this is the case, then, given 1, utilitarianism requires us to do things that, given 2, it requires us not to do. This would be incoherent.

The full argument could be laid out like this:

1. Having a project as a commitment
    a. is incompatible with regarding that project impartially or as dispensable;
    b. requires certain action robustly with respect to changes in rewards.
2. Utilitarianism requires us to
    a. regard all projects impartially and as dispensable;
    b. act in a way that is sensitive to rewards.
3. By 1 and 2, utilitarianism requires us not to have commitments.
4. Utilitarianism requires us to have commitments.

5. By 3 and 4, utilitarianism requires us both to have and not have commitments: therefore, utilitarianism is incoherent.

3.2. For premiss 4

How could 4 be supported? In one section of his *Critique*, Williams suggests that utilitarianism would be nonsensical if people did not have projects, of which commitments are a subset. He takes it 'that in talking of happiness or utility one is talking about people's desires and preferences and their getting what they want or prefer, rather than about some sensation of pleasure or happiness.'[8] (1973, 80). For utilitarianism to be meaningful, therefore, there must be preferences to satisfy. Williams assumes that this requires people to have projects from which those preferences arise. And this must include some projects that are not the utilitarian project itself, since conceived as a project of maximising preference-satisfaction, it is 'vacuous' unless there are 'other more basic or lower-order projects.' (1973, 110) Commitments are one class of those projects.

This argument fails to establish that utilitarianism requires us to have commitments. For one thing, the 'requirement' in question is more like a presupposition than a moral prescription. If utilitarianism would be vacuous without commitments, then the defender of utilitarianism might be glad that some people have them, for this makes her theory meaningful, but it does not follow that utilitarianism says that agents should adopt commitments. For another, even if we grant that utilitarianism requires us to have preferences and commitments are one class of thing that gives rise to preferences, it does not follow that utilitarianism requires us to have commitments. Imagine (with Parfit 2016, 118) a world whose inhabitants had only the drabbest of pleasures in their lives – muzak and potatoes. Suppose that they only have two preferences: that there should be muzak rather than silence, and potatoes rather than gruel. It is not that they have any deep affection for muzak or potatoes: they desire and enjoy them no more than we do. They simply lack the imagination to form preferences for anything else. They do not have commitments, in Williams's sense. Yet it is

[8] I prefer the term 'well-being' to 'happiness' or 'utility' to denote the end of utilitarianism, but I take it that these terms are interchangeable.

obvious what they prefer and therefore what utilitarianism recommends for this world: more muzak and potatoes.

Williams makes a better argument for 4, which stems from the observation that when people are happy, it is because 'they are involved in, or at least content with, something else.' (1973, 112) It is difficult to conceive of a person who is happy but not because of anything she is doing and enjoying. Well-being is not like toothache; a feeling that can come on or fade away. Typically, it involves some sort of activity – one is happy when and because one is eating, dancing, or succeeding in something. This view of well-being does not depend on the preference-satisfaction interpretation of well-being. A plausible hedonistic account would recognise that the mental states constituting well-being are either identical with or typically accompanied by these states of activity and enjoyment.

These states of activity and enjoyment are typically grounded in our projects: we do and enjoy things because of what we desire, value, pursue and identify with. Our well-being therefore has a lot to do with the projects we have, and since utilitarianism values well-being, it should require that agents take on the kinds of projects that make them happy.

But again, it does not follow from the fact that utilitarianism requires us to have projects that it requires us to have commitments. Williams proposes, as an empirical hypothesis, 'that many of those with commitments, who have really identified themselves with objects outside themselves, who are thoroughly involved with other persons, or institutions, or activities or causes, are actually happier than those whose projects and wants are not like that.' (1973, 113–14) Whether this hypothesis is true or not is an interesting and important question, and the answer is not obvious, although the prevailing view is that it is (see Calhoun 2009 for dissent). Raz writes that that our typical

> 'notion of a successful life is of a life well spent, of a life of achievement, of handicaps overcome, talents wisely used, of good judgment in the conduct of one's affairs, of warm and trusting relations with family and friends, stormy and enthusiastic involvement with other people, many hours spent having fun in good company, and so on.' (1986, 306)

Trusting familial relationships and friendships paradigmatically involve commitments, and achieving things, overcoming handicaps and wisely using one's talents may also do so. A life without these things might include good company and sound judgment (as well as more sensory pleasures that Raz fails to mention), but we might resist calling a truly happy life – or at least think that it would have been better with respect to well-being had it involved commitments. This tells in favour of Williams's hypothesis. On the other hand, some of the worst lives tend to involve commitments as well: loving marriages and friendships break up, dreams are unfulfilled, martyrs are made in defence of lost causes. It is likely that those with commitments *that are fulfilled* have happier lives than those without commitments, but this group is (tragically) only a subset of those with commitments. Nevertheless, I will grant that it seems likely there are some kinds of commitment such that for most of us, if we adopted them, our lives would be happier. If this is true then it is a reason for utilitarianism to recommend we adopt commitments, as premiss 4 says.

The case for 4 is strengthened when we also consider the effects our having commitments has on other people's lives. That relationships which increase the well-being of those who have them would be impossible without commitments means that my having commitments is not only good for my well-being, but good for the well-being of others too, insofar as they have the option of entering such a relationship with me. Without others in our community generally being committed to telling the truth, for example, it would be difficult for us to trust them, and trust is necessary to a range of social structures and interactions which are conducive to well-being. The typical commitment of parents to their children seems an efficient way of providing for children's well-being and their becoming considerate members of society, to the benefit of everyone. It is plausible that many of humanity's greatest achievements, which have had far-reaching effects on well-being, would not have been possible without commitments. Great discoveries and masterpieces are made by scientists and artists who are deeply committed to their work. Socio-political improvements such as the extension of human rights and public services, decolonisation and democracy were fought for – often literally – by people who built their lives around those causes.

On the other hand, it is difficult to know what a life, and still less the world, would look like in the absence of commitments. Perhaps people would be happier without the possibility of unfulfilled commitments. Perhaps such an uncommitted society would find new, looser kinds of relationships and social structures which would make people happy; perhaps it would provide socio-political goods without anyone having to fight for them. (There would probably be less fighting in general.) But the low resolution in which we can imagine such a world indicates the pervasiveness of commitments in ours. This pervasiveness supports the suggestion that commitments are a normal and perhaps necessary part of human life. Given the likely pains of a life that is far outside the norm for our species, it is probable that our well-being is greater if we have some commitments, and therefore, that utilitarianism requires us to do so.

3.3. The way ahead

To restate my reconstruction of Williams's argument:

1. Having a project as a commitment
    a. is incompatible with regarding that project impartially or as dispensable;
    b. requires certain action robustly with respect to changes in rewards.
2. Utilitarianism requires us to
    a. regard all projects impartially and as dispensable;
    b. act in a way that is sensitive to rewards.
3. By 1 and 2, utilitarianism requires us not to have commitments.
4. Utilitarianism requires us to have commitments.
5. By 3 and 4, utilitarianism requires us both to have and not have commitments: therefore, utilitarianism is incoherent.

In Chapter 2 I sketched Williams's argument for 1, 2 and 3. In section 3.2 I gave his justification for 4. I will not question the validity of the argument nor the suggestion that incoherence in the form described by 5 is sufficient reason to reject a moral theory. Therefore, to provide a response to Williams I must argue against at least one of this argument's premisses. My preferred response to the Integrity Objection focuses on premisses 1 and 4, arguing that insofar as one of them is true the other is false. That is, if commitments

are robustly demanding and incompatible with impartial consideration which countenances dispensing with them, they are not required by utilitarianism. This response appears in the final chapter, where I outline one way in which I think utilitarianism might, coherently and plausibly, advise us to live.

Another response to the Integrity Objection rejects premiss 2. It holds that utilitarianism does not require us to regard our commitments impartially and as dispensable, nor does it prevent robust action. Given the arguments of Chapter 2 this claim might seem impossible to justify. But this response makes a distinction between two things that I have hitherto conflated: the way that utilitarianism, as a theory, regards commitments and actions, and the way that it requires *us* to think and act. As this response holds that we are not required by utilitarianism to view the world as the theory itself does, it has been called 'self-effacing'. Though I will argue that there is a problem with the self-effacing response to Williams's objection, it provides important insights. I will now consider that response in greater detail.

4. The self-effacing response

4.1. Self-effacingness

The self-effacing response agrees with Williams that, as we saw in chapter 2, the utilitarian calculus is insensitive to the special relation between a person and their commitments, accounting for each in the commensurable currency of well-being and countenancing action at odds with them whenever the sums demand it. But it denies that, as 2a holds, utilitarianism requires *us* to regard our commitments in such a way, and that, as 2b holds, it would require *us* to perform actions at odds with them just because some potential reward renders those actions optimific.

In denying 2a, this response distinguishes two things: how utilitarianism looks at the world and how it requires agents to do so. Such distinctions, or similar ones, have a long history in utilitarian thought, and are found in Mill (2008, chap. 2) and Sidgwick (1962), as well as Williams's interlocutor J. J. C. Smart (1973, sec. 7). After Williams and partly in response to the Integrity Objection, it was used by Railton (1984) and Parfit (1984, chap. 1). (They defend consequentialism in general – Railton explicitly disavows utilitarianism on other grounds – but their remarks can be put to work defending utilitarianism.) Such distinctions, and the denial of 2, imply that utilitarianism does not necessarily require us to think about our choices through utility calculations; in fact, that it may require us *not* to think in the way the utilitarian calculus itself does. That it asks us to look at the world in a different way to how it does itself earns it the description 'self-effacing'.

How could utilitarianism require us to not consider our choices through impartial utility calculations alone? Recall our definition of utilitarianism: the moral theory that recommends that we choose the option that will maximise well-being. This is not the same as recommending that we choose the option that will maximise well-being *because it will maximise well-being*. Say that lending me some money will maximise well-being, because I will put it to more productive use than you would. Utilitarianism says that you should lend it to me. You have conformed to this recommendation whether you lend it to me out of a desire to maximise well-being or out of a desire to fulfil your duties of friendship. So we can do what utilitarianism requires without being motivated to do so by the factor that determines

what those requirements are (well-being, impartially considered). That is, utilitarianism need not ask us to *employ* utilitarianism as our decision-procedure.

Furthermore, our options can include not only actions, but also how we think. An option for an agent is any possibility that the agent is in a position to realise. Whilst the process is perhaps not as straightforward as realising action, we are able to realise thoughts, beliefs, attitudes and dispositions.[9] Utilitarianism says that we should do so in the way that maximises well-being. This way might not be equivalent to modelling one's thinking on a utilitarian calculus. In fact, such an equivalence is unlikely. Given such a non-equivalence, utilitarianism will recommend that we do not think in the manner of a utilitarian calculus – that we do not become 'subjective utilitarians'.

There are many reasons to think that this non-equivalence holds, and therefore that we should not, according to utilitarianism, be subjective utilitarians, employing the theory itself in our decision-making. Those that are specific to Williams's objection and the self-effacing response to it are investigated in greater detail in section 4.2. But to motivate the idea, consider these examples.

Time constraints can mean that it would be more conducive to well-being to act quickly, without spending time considering the consequences of one's options. Sometimes 'warm and spontaneous feeling' (Smart 1973, 45) is called for. It may be better to make more trivial choices out of habit or only considering circumscribed options and consequences, freeing up time to consider more important choices at greater length, making sure that one

[9] It may be objected that such realisations are, unlike the performance of actions, not objects of choice and therefore not options in the same way. I think that, as Pettit suggests, the options that agents can choose go beyond their actions: people do, to some extent, choose the way in which they think. It may be that this can only be realised through actions which cause changes in one's way of thinking – going to church more regularly, or reading a self-help book, for example. But the thing that the agent is choosing to realise is not simply that they go to church more, or finish the book, but that they become a more virtuous person, with thoughts, beliefs, attitudes and dispositions that they lacked at the start of the process.

chooses the correct utilitarian option in the latter.[10] Sometimes calculating the effects of different actions on general well-being can be done only extremely roughly if disaster is to be averted: acting from instinct or some quickly applicable rule would be preferable. Consider a driver who is heading towards a fallen pedestrian, and, in the next half a second, must slam on the brakes or swerve to avoid colliding with them. Asking herself which course of action will cause slightly lower carbon dioxide emissions, and therefore is more conducive to well-being is not a good idea. She should choose one of the two as quickly as possible.

Another area in which utilitarianism will not recommend employing utilitarianism is games. One could not successfully engage in any competitive game without suspending utilitarian decision-making. One must aim to win. Imagine, in football, a penalty-taker deciding where to put the ball not by asking herself how she would be most likely to score, but how she would be most likely to increase well-being. (If the opposition fans are fanatical enough, the answer to the second question might be to miss the goal entirely.) And imagine the goalkeeper, ball coming towards her, deciding whether to save it or not using a utilitarian decision-procedure. Perhaps making the save would upset the attacking team. Perhaps the game would be more enjoyable for spectators if a goal were scored. If these thoughts are dominant in their minds, then the penalty-taker and goalkeeper are not truly playing football. They are performing some sort of play aimed at eliciting certain responses and satisfying certain desires. If games are conducive to well-being – and millions of fans suggest that they are – then there must be a time, namely on the field of play, where utilitarian considerations are inappropriate, according to utilitarianism itself.

Sometimes employing utilitarianism would open our deliberation to biases that employing non-utilitarian rules would not. Imagine, plausibly, that when I try to work out whether a given act of theft would maximise well-being and therefore be justified for me to do, I am prone to 'cooking the books' towards an affirmative answer when my stealing would benefit me. I might conclude, employing utilitarianism with this bias, that in many cases

[10] According to Larissa MacFarquhar (2011), Derek Parfit wore the same outfit – white shirt, black trousers – every day, to save himself from having to deliberate about how to dress each morning.

I should steal – and in many of these I would be incorrect in utilitarian terms. Alternatively, if I eschewed utilitarian deliberation and followed the commandment 'thou shalt not steal', I would never steal, and so not make mistakes in these cases. (I might make mistakes in ones in which theft is in fact conducive to well-being – but might still be correct more often than I would be if I employed my biased utilitarianism.)

The need for quick shortcuts and for protection against bias make acting out of general dispositions and principles necessary. A disposition is a standing motivation to perform or refrain from some type of choice or act; a principle is a commandment such as 'thou shalt not steal'. These things can be applied to all choices, like utilitarianism, but unlike utilitarianism their application does not involve assessing the idiosyncrasies of each situation and performing calculations – rather, it is a matter of asking whether some option involves an act of the type covered by the principle, or acting more or less instinctively from the disposition. According to R. M. Hare:

> 'If it were not possible to form such dispositions, any kind of learning behaviour would be ruled out, and we should have to meet each new situation entirely unprepared, and perform an "existential" choice or a cost-benefit analysis on the spot.' (1981, 36)

Performing a cost-benefit analysis on the spot for every new decision is impossible – or at least impossible to do well – because we are bounded human agents, whose reasoning processes take up time and mental space (a point pressed by Williams (1981, 51–52)). Hare compares the agent with no intuitive principles to a person driving a car 'without having learnt to drive a car, or having totally forgotten everything that one had ever learnt – to drive it, that is, deciding *ab initio* at each moment what to do with the steering wheel, the brake, and other controls.' (1981, 36) Just as such a driver would often crash, the constant cost-benefit analyses of the subjective utilitarian would often go awry; not just through biases, but by the cognitive limits of human minds. Rather than accepting that we ought not steal, we would have to weigh up, in every shop, the expected utility of our paying for or shoplifting the item we wanted. Rather than accepting that one ought to keep promises, we would have to weigh up, when the moment came, whether it would satisfy more preferences for us to honour this particular promise or not. And next time we were in a shop, or had a promise to keep, we would be faced with a slightly different situation and with no general

principles would have to perform such an analysis from scratch again. We are more likely to choose the options that utilitarianism recommends if we act from more general principles than from utilitarianism itself – and this is not to mention the additional stress and opportunity cost of expending mental effort on so many cost-benefit analyses.

Sidgwick wrote:

> 'if experience shows that the general happiness will be more satisfactorily attained if men frequently act from other motives than pure universal philanthropy [i.e. do not employ utilitarianism as a decision-procedure], it is obvious that these other motives are to be preferred on Utilitarian principles.' (1962, 413)

The antecedent, given the considerations listed above, seems to be true. Therefore, utilitarianism recommends that we often do not employ utilitarianism, but deliberate in other ways.

The self-effacing response to the Integrity Objection makes a particular use of this thesis. It holds that given utilitarianism's recommendation that we do not always employ utilitarianism, it does not follow from William's observations about utilitarianism's view of the world that 2a is true. In fact, the response goes, 2a is false, since it would not be conducive to utilitarianism to regard all our commitments impartially and as dispensable. Instead, utilitarianism will recommend that we take the attitudes described in 1a for those commitments that, according to 4, it endorses. Thus, if there is, as Williams thinks, a tension between having commitments and thinking in a utilitarian way, this tension would result from our thinking in a way that utilitarianism warns us against and thus cannot cast any doubt upon it.

4.2. Impartiality and indispensability

To deny 2a, the self-effacing response must establish that utilitarianism does not recommend that we think of our commitments in the impartial way Williams describes, as 'one satisfaction among others'. It is plausible that this way of thinking would not be conducive to well-being. One reason for this is Williams's worry about alienation. If we regard our commitments in the detached and impartial manner of the utilitarian calculus, we become

somewhat alienated from them. As we saw in chapter 2, we come to see them as on a par with the commitments of others – even though *our* commitments are specially motivating for us and define our identity. Such alienation may be distressing, bringing forth painful feelings of dissociation and estrangement, and raising difficult questions about who we are. This suggests, ceteris paribus, that regarding our commitments impartially will not be conducive to well-being, and therefore that utilitarianism will not recommend this kind of thinking.

Another utilitarian consideration against treating our commitments impartially comes from Parfit, who observes:

> 'Most of our happiness comes from acting on certain strong desires. These include the desires that are involved in loving certain other people, the desire to work well, and most of the strong desires on which we act when we are not working.' (1984, 27)

The well-being that results from satisfying such desires can be fed into the utilitarian calculus. However, satisfying a desire and acting on it – that is, being motivated by it – are two different things. We do not act *on* those desires, whatever action results, if we deliberate in the manner of the utilitarian calculus: the grounds of our choice of action would be the likely effects on the general well-being, in which my desires play no bigger part, qua my desires, than anybody else's – just as Williams complains. If Parfit is right, then utilitarianism should require that we act on our strong desires, rather than on utilitarian grounds. And to act on our strong desires we must regard them from a partial point of view, treating them specially because they are ours. For we are aware that others have strong desires too, but we do not take these as grounds for action in the same way; if we did, we would be acting on the general sum of strong desires, not on *our* strong desires (and this is surely what Parfit means when he says that most of our happiness comes from acting on strong desires). Parfit's 'strong desires' seem to include Williamsian commitments – certainly, we act on commitments and often acting on them brings us happiness. Therefore, if Parfit's observation is correct, utilitarianism does not require us to regard our commitments in the detached and impartial way that the utilitarian calculus itself does, but rather to regard them as special sources of motivation for us.

Chapter 2 also raised Raz's idea of constitutive incommensurabilities (1986, chap. 13). There are goods, Raz claims, which we cannot enjoy unless we regard them as incommensurable with other goods. Friendship is claimed to be such a good: one cannot, in Raz's view, be a friend to someone if one considers one's relationship with them to be commensurable with money – that is, if it makes sense to you that there could be some sum of money of equal worth to that relationship. Utilitarianism makes all goods commensurable in the currency of well-being. A relationship is, according to utilitarianism, equivalent in worth to some sum of money, the sum that would bring about the same amount of well-being. If Raz is right, we cannot have friends if we think in this way. Friendship is uncontroversially conducive to well-being, and so utilitarianism should consider us to have friends. Therefore, utilitarianism encourages us not to regard our relationships as commensurable with other goods, even if on the level of theory it does so itself.

Are there utilitarian reasons not to view one's commitments as dispensable? One reason has to do with trust. Close personal relationships often depend on the parties to them believing that the other parties will stand by their commitments. For example, a marriage may become untenable when the partners cease to trust one another to keep their vows. If such relationships are conducive to well-being, then utilitarianism will recommend that we provide the conditions for such trust. In a close personal relationship, where the other party is likely to be able to perceive your internal monologue with some accuracy, it might undermine trust if that monologue were constantly reviewing one's commitment to the relationship, regarding it as dispensable if the occasion demands it. The best way to maintain the trust of one's partner, then, would be to regard one's commitments to them as indispensable. If the relationship is conducive to well-being, then so is maintaining this trust, and therefore utilitarianism will recommend considering one's commitments to be indispensable – even though utilitarianism itself constantly weighs them against other options by the currency of well-being.

In addition, treating some of one's projects as indispensable commitments could be part of a heuristic of the sort Hare noted the need for, helping to make decisions that are closer to the recommendations of utilitarianism than we would get if we treated them as dispensable. Above, I gave the

example of the commandment 'thou shalt not steal'. Following this plausibly generates more utility than not, because without this commandment a utilitarian agent might bias her decision-making in favour of stealing when doing so would benefit her at the greater expense of others. How should one treat that commandment? One could treat it as a mere guideline, to be set aside when the circumstances demand or permit. But doing so is likely to let bias creep back in – how does the agent judge when circumstances make it appropriate to set it aside? Why think she would not cook the books here as well? Alternatively, one could treat the commandment as an indispensable commitment, believing that it applies in all circumstances. Then one will never allow self-interested biases to lead to sub-optimific choices.

Treating less moralistic projects as indispensable may have similar benefits. Say that one wants to be a respectable philosopher. As Philippa Foot penetratingly observes, 'If one wants to be a respectable philosopher one should get up in the mornings and do some work, though just at that moment when one should do it the thought of being a respectable philosopher leaves one cold.' (1972, 306) If one regards one's desire to be a respectable philosopher as just another desire, which can be dispensed with in the light of other desires, one may easily talk oneself into staying in bed. This is true even if one would in fact better satisfy one's desires by getting up: the temptation to stay in bed, like the temptation to steal, biases one's deliberations so that one may become mistakenly convinced that doing so is for the best. However, if one regards being a respectable philosopher as an indispensable commitment, something that simply cannot be foregone, one will force oneself up and start working. Given that fulfilling such projects brings more well-being than staying in bed, in the long-term, utilitarianism will see this as a reason to treat them as indispensable commitments.

4.3. Robustness revisited

So far I have given reasons to think that utilitarianism – contrary to 2a - will not require that we consider our projects impartially or as dispensable. If it did, it would be incompatible with our having commitments, given 1a. What about 2b, which holds that utilitarianism requires us to modify our action

in a way that is sensitive to rewards, and hence fails to have the robustness demanded (according to 1b) by commitments? If 1b and 2b are both true, 3 would follow and granting premiss 4 utilitarianism would be incoherent, even if 2a is false. To deny 2b, the self-effacing response must deny that utilitarianism requires that, in all possible cases, we should perform the act which will bring about the greatest well-being. If this were accepted, then it would follow that if the reward made a positive difference to one's well-being any type of act could be recommended by utilitarianism, if the reward were big enough.

It may seem impossible for utilitarianism to deny this principle. After all, utilitarianism was defined by Mill as 'the creed which accepts as the foundation of morals… that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.' (2008, 137) However, utilitarianism as I am defining it here ranges not over actions, but *options*. An option is 'a possibility which the agent is in a position to realise or not' (Pettit 1993, 232). One common option is the agent's performance of an action. But as we saw in the previous section, there may be other kinds of option, such as how the agent thinks. Sometimes the options available to an agent are sets of actions (Parfit 1984, chap. 1) or the adoption and maintenance of commitments.

This can be brought to bear by the self-effacing response on the question of robust action with respect to commitments. The problem is that the utilitarian agent would 'sell out their commitments' for a big enough reward. Consider George, in a variation of his case in which the rival chemist would pursue the research no more dangerously than he would, but George is offered a financial bonus (which his competitor would not be offered) to take the job. If this bonus is large enough, taking the job might be more conducive to well-being than refusing it (i.e. it is the optimific action). 1b implies that if George would take the job for the bonus, he could not have a commitment to opposing chemical warfare, and if he had the commitment he would be unmoved by the bonus. Utilitarianism seems sensitive to rewards in a way that precludes commitment.

However, the utilitarian might respond, if George's commitment to opposing chemical warfare makes him incapable of accepting the job then once he has this commitment doing so is not an option for him. If it is 'absurd' for utilitarianism to demand it of him, because he cannot do it, then

utilitarianism will not ask him to. Ought implies can: morality ranges only over things that we can do or not do.

It does not seem right, though, to regard George's refusal of the job as a matter of fate. It is a voluntary act. It is only impossible for him to accept the job (according to 1b) given that he has this commitment. But if his commitments were otherwise he could do otherwise, and adopting and maintaining a commitment is something an agent can realise or not, and therefore can be an option for them.

Grant that if George has the commitment, he will refuse the job. (If this conditional is false then 1b is, because if it is possible for a committed person to take the job for some reward, they do not act robustly with respect to their commitment.) Either he can get rid of his commitment, or he cannot. If he cannot, then he cannot but refuse the job; acceptance is not an option for him and therefore utilitarianism does not direct him to accept. Furthermore, if at some previous point in his life George *did* have a choice about whether to adopt this commitment or not, it is plausible that utilitarianism would have required him to do so. In this case, utilitarianism says that George should have his commitment to opposing chemical warfare, and if this involves refusing the job, then there is a sense in which utilitarianism says that he should refuse it – even though this action is not optimific. If, on the other hand, George is able to dispense with his commitment, accepting the job is an option for him. In fact, accepting the job is an option for him only if he can drop his commitment, if it is true that having such a commitment necessitates refusing such jobs. But even in this case, utilitarianism might not ask George to accept the job, even if the reward offered makes doing so optimific. George's options are not simply *accept the job* and *reject the job*; instead, they are *accept the job + drop the commitment* and *reject the job + retain the commitment* (he also has the option of refusing the job and dropping the commitment, but this is insignificant). Now, if premiss 4 is true, then utilitarianism requires us to have some commitments. This could only be because having those commitments is more conducive to well-being than not having them. If George's commitment to opposing chemical warfare is one such commitment then retaining the commitment has positive weight in the utility calculation. Given this weight, *reject the job + retain the commitment* could rank higher than *accept the job + drop the commitment* even though

accepting the job would be more conducive to well-being than rejecting it. This would be true if the difference in effect on well-being between having and not having the commitment were greater than the difference between accepting and refusing the job.

Therefore, utilitarianism need not recommend that we act against our commitments just because there is some reward on offer that makes such action optimific. If the action is truly impossible for us, utilitarianism does not recommend it (and may even recommend adopting the commitment that makes it impossible). If the action is possible for us, and it is not possible to have the commitment and perform the action, then it is because it is possible for us to dispense with the commitment in question. If so, we face the options of retaining the commitment and not performing the action or performing the action and dropping the commitment. As long as having the commitment is conducive to well-being, utilitarianism may recommend that we do not perform the action.

The idea behind 1b's claim that commitments are robustly demanding is that one is not really committed to some cause if one *would* perform an action that is at odds with it simply for some reward to oneself (whatever one actually *does* do). The argument just given demonstrates that utilitarianism need not ask us to perform such actions where well-being conducive rewards are on offer. Thus, the self-effacing response can say, when we act in accordance with those commitments, that action can be appropriately robust even if we are utilitarians, because we would still perform it if we were offered a reward to do otherwise.

Is this too fast? Perhaps. What it establishes is that utilitarianism can recommend actions that are not themselves the most conducive to well-being, when they form part of an option which involves retaining a commitment. This is true where dropping the commitment is impossible, because performing the action is not an option for the agent. It is also true in cases where the commitment can be dropped because commitments can have value to utilitarianism. But this value stems from their effects on well-being. And this makes commitments commensurable with other goods. As long as utilitarianism views commitments and actions as commensurable, there will be some size of reward that it will recommend we take even if it means breaching our commitments.

Return to the case of George. His options, we decided, are not *accept the job* and *reject the job*; instead, they are *accept the job + drop the commitment* and *reject the job + retain the commitment*. As long as maintaining the commitment makes some contribution to well-being, it is possible for the latter to be the option utilitarianism recommends, even if, as isolated actions, accepting would lead to more well-being than rejecting. However, it is also true that if the difference in terms of well-being between accepting and rejecting were greater than the difference between dropping and retaining the commitment, utilitarianism would recommend accepting the job and dropping the commitment. The former difference could be made great enough, in principle, by the offer of a very large reward. So all that the move from actions to options and recognition of the utilitarian value of commitments does is raise the size of the reward needed to make utilitarianism advise George to take the job.

The question is whether this endangers the robustness commitments require. Robustness comes in degrees. Take Shakespeare's line that 'Love is not love, that alters when it alteration finds'. It is plausible that if one loves someone, it cannot be the case that if they were to become poorer, or fall ill, or change the colour of their hair one's feelings for them would change. Love is robustly demanding across these possibilities. But it is not with respect to some other alterations. If, counterfactually, one's beloved were to fall in love with someone else, or if their character changed drastically, or they started making much greater demands on you, then one might adjust one's attitude towards them – and this would not mean that, as things stand, one does not love them. The world is littered with people who have truly loved others and stopped. So robustness need not be absolute; it would be strange for Williams to demand that having a commitment requires that there are no possibilities in which one would act at odds with it.

Indeed, this is not what he demands. As I have reconstructed his argument, having a commitment only requires that one's actions in respect of it are robust against rewards: one could not be tempted to act at odds with some commitment by an offer of personal benefits. I have given, on behalf of the self-effacing response, an argument that utilitarianism does not ask us to act at odds with our commitments in every case in which a reward is on offer which is big enough to make that action the most conducive action to well-being. This is because the action may be part of an option that is not the

option most conducive to well-being. If this argument succeeds, then the action of a utilitarian agent with respect to her commitments may be robust across the set of possibilities that involve her being offered some reward under a certain threshold, even if that reward made the action the most conducive to well-being. But it would not be robust across possibilities that involve rewards bigger than that – big enough to make the option, of which that action is a part, the most conducive to well-being.

Is robustness above this threshold necessary for having commitments? The answer depends on where the threshold is. If it is low, so that a small reward would trigger action at odds with some alleged commitment, then it would be hard to claim that the agent was truly committed. If it is high, so that only a very large reward would trigger such action, then a lack of robustness above the threshold may not threaten the claim to have a commitment at all. There are reasons to think that a high threshold will follow from utilitarianism.

The first is the disutility to the agent of breaching their commitments. As Railton puts it: 'Commitments to others or to causes as such may be very closely linked to the self, and a hedonist who knows what he's about will not be one who turns on his self at the slightest provocation.' (1984, 142) One of the things that gives us happiness, or is a condition for it, is a secure sense of identity. The questioning of one's identity – being told that one is not a real philosopher, learning that one is adopted, being misgendered – is distressing. In failing to honour a commitment, one calls one's own identity into question, which may be similarly painful. Examples are people breaking the rules of their religion or betraying those they love: such failures are usually accompanied by a painful guilt. If utilitarianism is to recommend actions which break with commitments, the payoff for doing so, in terms of well-being, must be high enough to outweigh this loss of well-being.

A second reason that utilitarianism will require very high rewards before it asks us to act against our commitments is that commitments determine repeated actions across a long period time. If I have some commitment – say, to a spouse – then I will do things to help her in her projects, give her pleasure, lessen her burdens, and so on, *daily*. If such a commitment is recommended by utilitarianism (and it is only those that are that interest us, given Williams's charge of incoherence) then that is presumably because

having it makes me more likely to do these things and these things are conducive to well-being. Having the commitment, therefore, means that a number of actions conducive to well-being are performed over a long period of time: the sum of well-being generated by that commitment is therefore likely to be very large. If one were to lose the commitment for the sake of performing just one action, on utilitarian grounds, then that action must have an even larger positive impact on well-being, outweighing the sum of all the possible actions dependent on the commitment. If a reward being attached to that action is what gives it such an impact, therefore, the reward must be very big.[11]

If these considerations establish that the reward that would be necessary in order for a utilitarian agent to breach her commitments is very large, is this enough to deny 2b? That premiss states that utilitarianism requires us to modify our actions in a way that is sensitive to rewards. Strictly, these remarks do not refute that. However, sensitivity comes in degrees. What I hope to have shown is that utilitarianism does not require us to modify our actions in a way that is *highly* sensitive to rewards, even if it entails that there is some conceivable reward so great that it should move us to breach our commitments. This delivers a greater robustness to the actions of a utilitarian agent than might be expected.

4.4. A criterion of right action?

Self-effacing utilitarianism recommends that we sometimes perform actions that will not maximise well-being, knowing that they will not maximise well-being. This is true, for example, in the modified case of George discussed in the previous section, in which refusing the job is less conducive to utility than accepting it, but is a necessary part of a utility-maximising option (which includes his having a commitment to oppose

[11] In fact, it could be that, for some commitments, the reward that would induce a utilitarian agent to break them would be impossibly big: if there is an upper limit to how much well-being a reward to one person can bring, and the utility attached to having the commitment is greater than this limit. The diminishing marginal utility of most goods makes it plausible (though does not entail) that there is some such limit.

chemical warfare). When he receives the offer, George already has the commitment, and is incapable of keeping it if he accepts the job. So self-effacing utilitarianism doesn't ask him to accept, despite it being true that doing so would be optimific, taken as an individual action. This would also be true if George were unable to rid himself of his commitment, even if it were not conducive to well-being, since accepting the job would be impossible for him.

It might be objected that the self-effacing response falls into the very trap that it aims to avoid: incoherence. Williams wrote of utilitarian responses that aim to affirm the value of commitments that

> 'The difficulty is that such dispositions are patterns of motivation, feeling and action, and one cannot have both the world containing these dispositions, and its actions regularly fulfilling the requirements of utilitarianism.' (1981, 51)

The idea is that given commitments, agents will perform non-optimific actions. But utilitarianism says that we should always do the optimific thing, so cannot recommend these actions, and therefore cannot recommend commitments.

This objection is not successful. It trades on the conflation of actions and options that include actions. Utilitarianism, as I have defined it, asks us to choose the options that are conducive to well-being. If it is true that George's commitment necessitates his refusing the job, then he can only accept the job if he drops the commitment. *Ex hypothesi*, refusing the job and retaining the commitment is a better option, in terms of well-being, than accepting the job and dropping the commitment. So utilitarianism will recommend that option – and it will therefore not say that George should accept the job. Cases like George's do not show incoherence; they show that, according to utilitarianism, actions that are not optimific are not always thereby actions that we should not perform.

There is a significant difficulty for the self-effacing response in this neighbourhood, however. I have been careful so far to talk about the outputs of utilitarianism as recommendations and judgments on what we *should* do. Utilitarianism is often characterised differently: as a *criterion of right action*. (We have already seen Mill's definition, and Wiggins's.) Proponents of the self-effacing response sometimes define their view as the separation

of utilitarianism as a criterion of rightness and as a decision-procedure. Railton writes that he holds 'the view that the criterion of the rightness of an act or course of action is whether it in fact would most promote the good of those acts available to the agent.' (1984, 152) Specified for utilitarianism, this would yield the following criterion of rightness:

> An action is right if and only if it is the most conducive to utility of all available actions, and wrong otherwise.[12]

Parfit has a slightly different view (1984, 24–25). He distinguishes what is objectively and subjectively right and wrong. Objective rightness is unaffected by the epistemic position of the agent: the criterion of rightness written above is a criterion of objective rightness. A criterion of subjective rightness (or wrongness) determines which actions are right and wrong given what the agent believes, or ought to believe. It is subjective wrongness that gives us grounds for blame. Parfit says that consequentialism claims that 'If someone does what he believes will make the outcome worse, he is acting wrongly.' (1984, 24) Following this line of thought, a utilitarian criterion of subjective wrongness would look like this:

> If someone performs an action that they (ought to) believe is not the available action most conducive to utility, they are acting wrongly.

According to either of these criteria, it would be wrong for George to refuse the job, since taking the job would be optimific, and he (we can stipulate) is in a position to know this. Together with the claim that utilitarianism says that George in some circumstances should refuse the job, we have a strange-sounding result: according to self-effacing utilitarianism, it can be the case that one should do something that is wrong.

Railton and Parfit embrace this result. Railton writes, of a case in which Juan spends time travelling to see his wife Linda when it would have been better if he had used that time to raise money for charity:

> 'The objective act-consequentialist will say that Juan performed the wrong act on this occasion. Yet he may also say that if Juan had had a character that would have led him to perform the better act, he

[12] This combines Railton's stated view with the assumption that all actions that are not right are wrong. This may be denied, but as Railton's description of the Juan and Linda case (below) shows, Railton himself seems comfortable with it.

> would have been less devoted to Linda… Thus it may be that Juan should have (should develop, encourage and so on) a character such that he sometimes knowingly and deliberately acts contrary to his consequentialist duty.' (1984, 159)

Parfit writes, of a case in which Clare saves her child at the cost of the lives of several strangers, from her maternal commitment:

> 'Clare could say: "I act wrongly because I love my child. But it would be wrong to cause myself to lose this love. This bad effect is part of a set of effects that are, on the whole, one of the best possible sets of effects. It would be wrong for me to change my motives so that I would not in future act in this kind of way. Since this is so, when I do act wrongly in this way, I need not regard myself as morally bad… There can be moral immorality, or blameless wrongdoing."' (1984, 32)

Because George, Juan and Clare should have their commitments (to opposing chemical warfare, to Linda, and to her children respectively), and these commitments necessitate wrong actions, they should perform wrong actions. This result is strange, but does it necessarily show that something has gone awry? Not all moral attitudes must align: some people find it hard to distinguish the view that one should not sell cannabis from the view that doing so should be illegal to do so. They are wrong. But could we distinguish what one should do from what is right and wrong? If I claimed that some act is wrong but that this doesn't imply that you shouldn't do it (and that in fact you should) you might question whether I was using 'wrong' in a meaningful manner. As Williams put it:

> 'If a man has a disposition of a kind which it is good that he has, and if what he did was just what a man with such a disposition would be bound to do in such a case, but (as I claim must sometimes be so) was counter-utilitarian: what is the force of saying that what he did was as a matter of fact wrong?' (1981, 53)

Williams notes that it cannot mean that if he had deliberated better, he would have done otherwise: according to self-effacing utilitarianism, one can deliberate in the best possible way and come to a sub-optimific conclusion, as in the cases of George, Juan and Clare. Neither can it mean that we ought to bring our children up to be the sorts of people who do

otherwise: according to self-effacing utilitarianism, one can be the best possible sort of person and come to a sub-optimific conclusion.

Mill wrote:

> 'We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience.' (2008, 184)

This provides some way of giving force to the judgment that X-ing is wrong that doesn't depend on any relation to the judgment that we should not X (or would not X if we deliberated better, or should bring children up to refrain from X-ing). However, it is not an attractive route for those endorsing self-effacing utilitarianism and the utilitarian criterion of rightness. According to the utilitarian criterion of rightness, actions that are not optimific are wrong. But not all non-optimific actions are such that utilitarianism says that we should punish them. This is especially so if we consider the cases of Clare and Juan. They perform non-optimific actions, called 'wrong' by Parfit and Railton. But it is absurd to think they should be punished. Rather, as Parfit says, they should not even be blamed.[13] Remember that utilitarianism says that we should do what will be most conducive to well-being, and it is unlikely that punishment in these cases would be: it would discourage the commitments and character from which they acted, and from which self-effacing utilitarianism endorses acting.

So if self-effacing utilitarianism is to adopt the utilitarian criterion of rightness, it has a problem in giving force to the judgments of that criterion. Is this a problem for the self-effacing response? Not necessarily; it is just a problem for its conjunction with that criterion. The self-effacing response is a defence of utilitarianism as I defined it in chapter 1, which is not committed to the utilitarian criterion of right action, but to utilitarianism as a criterion for determining which options we should choose. One response is to subsume talk of right and wrong into the question of what we should

[13] Stephen Darwall claims that 'What is wrong is what would be blameworthy were it to be done without adequate excuse.' (2010, 263) That will be problematic for self-effacing utilitarianism to accommodate for the same reasons Mill's principle is. (Unless 'adequate excuse' covers all cases in which blaming the wrongdoing agent would not be optimific.)

do, so that right actions are defined as those that are part of the options favoured by utilitarianism. Another is to omit talk of right and wrong altogether, as Norcross (2006) advocates. Another would be to find a different force for judgments of right and wrong issued by the utilitarian criterion of rightness. I leave the question open. In this thesis I talk about the recommendations of utilitarianism in terms of what we should do, and the important thing to note is that this need not translate seamlessly into talk of what is right: it is possible that utilitarianism says that we should do something, and utilitarianism is true, and doing that thing is not right.

4.5. What the response achieves

Recall my reconstruction of Williams's argument:

1. Having a project as a commitment
    a. is incompatible with regarding that project impartially or as dispensable;
    b. requires certain action robustly with respect to changes in rewards.
2. Utilitarianism requires us to
    a. regard all projects impartially and as dispensable;
    b. act in a way that is sensitive to rewards.
3. By 1 and 2, utilitarianism requires us not to have commitments.
4. Utilitarianism requires us to have commitments.
5. By 3 and 4, utilitarianism requires us both to have and not have commitments: therefore, utilitarianism is incoherent.

As we saw in chapter 2, utilitarianism treats our commitments in ways that we, if we have commitments in the way Williams describes, cannot. The self-effacing response points out that utilitarianism asks us regard our commitments in a utilitarianism manner only if doing so is conducive to well-being, and it is likely that such treatment is not. This enables a denial of 2a: utilitarianism does not require us to regard our commitments impartially or as dispensable. Given Williamsian claims about the necessitation of action by commitments, utilitarianism does not even recommend that we act contrary to them whenever doing so would be the act most conducive to well-being – its sensitivity to rewards is less than it first appears, at least mitigating 2b.

If the self-effacing response makes good on its claims, and I have so far given considerations to suggest that it does, premiss 2 is false. If 2 is false, then 3 is false; if 3 is false, then 5 is false. In this case there is no Integrity Objection to utilitarianism. Instead, there is an observation that utilitarianism requires us to have some commitments, and if 1 is correct this involves a certain way of being, thinking and acting which *prima facie* seems to sit awkwardly with utilitarianism. If utilitarianism requires such a committed life, this would be an interesting and significant finding.

5. Problems for the self-effacing response

5.1. Partly or wholly self-effacing?

Utilitarianism would be wholly self-effacing if it recommended that we never think in a utilitarian manner. It would be partly self-effacing if it recommended that we sometimes think in a non-utilitarian manner, and sometimes as subjective utilitarians. All plausible versions of utilitarianism are at least partly self-effacing: for the reasons given in 4.1, it is highly unlikely that it would maximise well-being if we employed the utilitarian calculus in all cases.

How far does the self-effacing response need to go in order to respond to Williams? In my reconstruction of the Integrity Objection, premiss 1 insists that having commitments is incompatible with regarding those commitments impartially or as dispensable. Does this mean that one must never regard one's commitments in that way, or merely must not always do so? Williams's answer is clearly the former. When he asks (rhetorically) 'how can a man, as a utilitarian agent, come to regard as one satisfaction among others, and a dispensable one, a project or attitude round which he has built his life?', the claim is that a man could never do this, not that he could not always do it: 'sometimes' would not be an appropriate answer. This does not mean that a wholly self-effacing response is necessary to respond to Williams, however. If utilitarianism said that the utilitarian way of thinking could be employed in cases where commitments are not involved, it would be partly self-effacing, but still never ask us to regard our commitments impartially or as dispensable.[14] Call such a view 'commitment self-effacing'. Given the importance of commitments to our moral lives (which is affirmed by the self-effacing response) such a view would be, though only partly, still very significantly self-effacing.

The self-effacing response, then, can claim that utilitarianism is either wholly or only commitment self-effacing. In this chapter I will focus on the

[14] In the final chapter, I sketch a partly self-effacing utilitarianism that does ask us to sometimes consider our commitments in the utilitarian manner and thereby falls foul of premiss 1.

claim that it is wholly self-effacing. Much of what I say, though, also goes for commitment self-effacing utilitarianism, within the context of cases involving commitments, which is a context of the utmost importance. The claim, for example, that utilitarianism could not guide us if it were self-effacing should be read as the claim that a wholly self-effacing utilitarianism couldn't guide us, and commitment self-effacing utilitarianism could not guide us in cases involving commitments.[15]

Considering the self-effacing response to his objection, Williams concluded that it showed that 'utilitarianism's fate is to usher itself from the scene… and utilitarianism has to vanish from making any distinctive mark on the world' (1973, 134–35). He considered that this would be a significant cost to utilitarianism as a moral theory, and I agree. In the next three sections, I develop three different problems for self-effacing utilitarianism that could come under the broader charge of 'ushering utilitarianism from the scene'. The first has to do with the guidance utilitarianism could give to an agent, the second with the force of its judgments, and the third with our belief in it.

5.2. The guidance problem

One function of a moral theory is to help agents to answer the question: 'what should I do?' This is why if utilitarianism were incoherent in the way Williams's objection alleges, that would be a severe problem for it. A theory that says both that one should and should not have commitments does not give a useful answer to this question regarding an important part of one's life. It may be objected that self-effacing utilitarianism cannot fulfil this function. According to the self-effacing response, utilitarianism does not require agents to reason as the utilitarian calculus does – it permits non-

[15] The writers I have associated most with the self-effacing response, Parfit and Railton, do not advocate wholly self-effacing consequentialism. They believe it is unlikely that the best consequences will come of us never considering things in a consequentialist manner (Railton 1984, 155; Parfit 1984, sec. 17). But they both say that if they are wrong about this, and consequentialism is wholly self-effacing, that would not tell against it. I disagree with this latter claim, but in chapter 6 develop a partly self-effacing view that draws on some of their remarks defending the former.

utilitarian motivations, such as those involved in commitments. When a subjective utilitarian asks themselves 'what should I do?', they reach the answer by identifying each option available to them and predicting and evaluating their consequences. If we are not to be subjective utilitarians, as self-effacing utilitarianism recommends, how should we reach our answer? It depends entirely on which way of reaching our answer would maximise well-being. Williams writes that

> 'If utilitarianism indeed gets to this point, [then it] determines nothing of how thought in the world is conducted, demanding merely that the way in which it is conducted must be for the best' (1973, 135)

At this point utilitarianism would not fulfil a necessary function of any moral theory: guiding our deliberation about what we should do. This is the guidance problem.

One response that could be given on behalf of self-effacing utilitarianism is that it does give an answer to the question 'what should I do?' – as a form of utilitarianism, it replies with whatever is the option most conducive to well-being. It is true that for any choice, self-effacing utilitarianism holds that one of the options, the one that maximises well-being, is the one that should be chosen. In this regard it is like subjective utilitarianism. The distinguishing feature of self-effacing utilitarianism is that it does not demand a utilitarian method of arriving at an answer, but this does not stop it giving one. This response fails because the guidance a moral theory must provide is not merely a matter of spitting out an answer to a question. It is rather a matter of making clear the factors that the agent should consider, and how she should reason in order to reach this answer. If I carried a black box around with me, which observed all the morally relevant features of my circumstances and told me what I should do when I asked it, but told me nothing of how it reached its outputs or what its inputs were, then it would not be right to say that the box guided me as a moral theory does. That is what self-effacing utilitarianism seems to offer when it tells us what we should do. Subjective utilitarianism offers more: it asks us to consider the features of each option that could affect well-being, to predict the likely outcomes of our actions with respect to them and to choose the option which will be most conducive to well-being. Though it will arrive at the same

answer as the self-effacing utilitarian black box, it does so with transparency about how it works.

So the charge is that self-effacing utilitarianism cannot provide the sort of guidance that not only answers the question 'what should I do?' by selecting some option, but helps us to reason our way to that option. Williams claim is that self-effacing utilitarianism 'determines nothing of how *thought* in the world is conducted' (my emphasis). Whilst it is clear how a subjective utilitarian thinks, it is not clear how an agent should go about their deliberations according to self-effacing utilitarianism. According to the self-effacing response, agents should sometimes act out of commitments; that can involve using as a premiss or justification that one has some commitment or other (such as when George says that he cannot accept the job '*since* he is opposed to chemical and biological warfare' (Williams 1973, 98 - my emphasis), or 'just acting' (1973, 118) in accordance with the dispositions associated with the commitment. But sometimes they should be motivated in other ways: their deliberation should happen however would be most conducive to well-being.

Parfit says that Williams's charge that self-effacing utilitarianism would therefore determine nothing about how thought is conducted

> 'is puzzling since, as Williams also claims, [utilitarianism, were it self-effacing] would be demanding that the way in which we think 'must be for the best'. This is demanding something fairly specific, and wholly Consequentialist.' (1984, 42)

In Parfit's view, self-effacing utilitarianism does determine how we should think. In fact, it determines it uniquely: we should think in the way, whatever it is, that maximises well-being. If this means acting from commitments, then self-effacing utilitarianism has determined that we act from commitments. If acting from purely utilitarian motives is not conducive to well-being, then self-effacing utilitarianism that tells us not to think in this way. Utilitarianism still *determines* how we think, even if it is not itself how we think. And this determination is done purely on the grounds of effects on well-being. Williams's claim is untrue.

However, a version of the black box problem applies on this level as well. Self-effacing utilitarianism may say that we should deliberate with commitments, and not like a subjective utilitarian, but it cannot be

transparent about how it comes to that judgment. Imagine that an agent asks themselves, 'how should I think about this choice?'. They may consult a utilitarian black box that says 'by understanding your commitments and sticking by them'. This may be enough to say, with Parfit, that self-effacing utilitarianism provides a determinate and specific answer. But the reflective person who poses questions about how they should think does not simply want an answer, they want to be shown how to reach an answer, or at least what justifies the answer that is given. Now self-effacing utilitarianism faces a dilemma. On one horn, it can refuse to give that kind of response, insisting that a standalone answer is sufficient. It could be argued that since self-effacing utilitarianism has provided an answer of what to do and how to deliberate, it fulfils the guidance function. But it would not provide the right kind of guidance about the higher-order question of how to deliberate. And this should be important to utilitarians, because, as the self-effacing response itself emphasises, the way in which we deliberate has significant effects on well-being. To give an unsatisfying answer to such an important question is a failure: it means that agents cannot justify to themselves or others why they should think as they do. Furthermore, we do not have black boxes telling us how to deliberate in different situations. To work out our own answers to such questions, we need some further general principles. Without these, we would be in a worse position, with respect to the question of how we deliberate, than Hare's agent trying to work out, in every situation and from a blank slate, which option would maximise well-being. We would not even know that we should aim at the decision-procedure that is most conducive to well-being. We would be totally without guidance, even though there is a theory out there (though not one that we are moved by) that gives determinate answers to the question of how we should think.

On the other horn of the dilemma, self-effacing utilitarianism can open up the black box, showing the agent its workings – that is, that the reason one should act from one's commitments is that this is best for well-being. Then we would be in a position to deliberate our way to selecting the right decision-procedure. But according to Williams, as we saw in chapter 2, if the agent accepts this justification, she cannot continue to have commitments. I will develop this line of thinking in the next chapter: but note for now that it is only partly, and not wholly, self-effacing (and not commitment self-effacing either): it does require that in some way, in

deliberations related to commitments, the agent is motivated by utilitarianism itself.

5.3. The force problem

Utilitarianism, including in its self-effacing form, is a theory that says what we should and should not do. It is commonly thought that such judgments have some typical or necessary connection to other attitudes. When we learn that we should (not) do something, this fact encourages us to adjust our stance towards that thing: it would be odd to say, 'you should X, but don't let that affect how you feel towards X-ing'. The judgment that we should do something should lead to a motivation to do that thing and dispositions to blame those who fail to do it and to advise others to do it. If a moral theory cannot explain, or attenuates, the relationship between the judgment that we should do something and these attitudes, that is grounds to doubt whether it really tells us what we should and should not do: such judgments are ones with the force to provoke further responses. Self-effacing utilitarianism seems to be in this position. This is the force problem.

Take motivation first. Learning that one should X typically motivates one to X, *ceteris paribus*. Consider the subjective utilitarian: if they learn that utilitarianism says that they should X, this will push them towards X-ing. This is not simply out of blind allegiance to the theory, a motivation to do whatever utilitarianism says because utilitarianism says it. It is rather because that utilitarianism says that one should X entails that X-ing is the option that would be most conducive to well-being for one to choose, and they are motivated to choose the options that would be most conducive to well-being because they are most conducive to well-being.

Now, self-effacing utilitarianism retains the idea that the judgment 'one should X' entails that X-ing is the option that would be most conducive to well-being. However, it differs from subjective utilitarianism in recommending that one is *not* motivated by a concern to maximise well-being. So the thing that makes it the case that one should X should not motivate one to X. Therefore, the self-effacing utilitarian cannot explain why learning that one should X typically motivates one to X, if one lives as

self-effacing utilitarianism itself recommends – at least, it cannot explain this in the same way that subjective utilitarianism does.

Nor should self-effacing utilitarianism appeal to blind allegiance to its moral judgments. Consider this response: 'Learning that one should X will motivate agents to X just because they are committed to doing as they should. There is no need for them to be motivated by the thing underpinning or entailed by the judgment that they should X.' Certainly, we do sometimes 'act from the motive of duty' doing things just because we believe they should be done. Children typically act on such grounds, refraining from doing something because it is naughty, trying to be good, and so on. But reflective agents will sometimes ask not only what one should do, but why it is that one should do it. And the answer to this question should be capable of motivating them. Not only this, but if all our motivation to act morally came from the motive of duty itself, Williamsian worries about compatibility with commitments would return. If we show care towards our friends and family just because we think we should (especially, perhaps, if we do not know what grounds this judgment), this would be somewhat alienated compared with the motivation to do these things for the sake of our friends and family themselves.

A better response for self-effacing utilitarianism to make is that learning that one should X should motivate one to X, but neither for the sake of well-being or the sake of doing what one should. Rather, the agent believes that the judgment that one should X entails something other than that X is conducive to well-being, and it is this other fact that motivates them. So, the agent may believe that 'one should X' entails that X-ing is a requirement of some moral virtue they aspire to, or some commitment they have. But this gives self-effacing utilitarianism a dilemma. Either some non-utilitarian motivating fact is always true of optimific options, or not. If so, whenever that non-utilitarian motivating fact is true of an option, one should choose that option, and one should be motivated by that fact to do so. In this case, both the criterion of what one should do and the motivation to do it are supplied by some non-utilitarian fact – even if it is true that one should always do the thing that utilitarianism recommends, it seems that utilitarianism would then be superfluous, as one should also always do the thing that some other theory recommends, and furthermore that theory does something utilitarianism cannot, which is motivate. Utilitarianism

would seem to be entirely ushered from the scene. On the other hand, if it were not the case that some non-utilitarian motivating fact is always true of the options most conducive to well-being, then the agent could only be motivated to do what they should by false beliefs. They are motivated by the belief that a certain fact is true of the option, and this underpins the judgment that they should do it. For some of the options they should choose, this fact will not be true.

Some people may not be worried by this. Some false beliefs can be good for us: a dose of optimism may even be necessary to keep us going, even if optimism means thinking that things will turn out better than they in fact will. But there is a problem about how these beliefs are generated. The self-effacing utilitarian endorses the idea that there are things that we should do, but that we should be mistaken about why we should do them. How then do we come to our judgments about what we should do? Presumably we should ask which options instantiate the fact that we falsely believe to underpin our moral judgments. But then there is no guarantee that we will tend to get these judgments right – in fact, if we correctly identify which options instantiate this fact we will get some of the judgments wrong. And then we would not conform to utilitarianism. One solution is what Williams called 'Government House utilitarianism' (1988, 188), calling to mind a colonial elite who use utilitarianism to make the judgments, which they promulgate to their subjects who are deceived about the basis of those judgments and don't try to make any judgments themselves. This does not seem a welcome implication of utilitarianism. At the very least, if utilitarian moral judgments are to have motivating force, for the self-effacing utilitarian who does not want to usher utilitarianism entirely from the scene, there are prima facie costs stemming from the fact that agents must be mistaken about morality.

Even if self-effacing utilitarianism makes it hard to account for the connection between the judgment that one should X and one's motivation to X, there are other forces such judgments are typically taken to have. For example, we tend to blame people who should X and fail to (without excuse), and to advise people who should X to X. These connections also create significant problems for self-effacing utilitarianism.

Blaming someone for something, or advising them to do something, are options – and therefore according to utilitarianism we should do these

things if and only if they are conducive to well-being. Now, there are cases in which someone should X, according to self-effacing utilitarianism, although it would not be conducive to well-being to blame them for failing to do so. Blame can have positive effects when it acts to incentivise or inform better behaviour (Wallace 2017), but it always involves bad feeling on both the part of the blamer and the blamed, and so in many cases makes a net negative contribution to well-being. For example, consider the following case. Your friend Mario, angry at his company failing to win a lucrative arms contract, deliberately stamps on a child's foot in a busy restaurant at which you are having lunch with him. You and the child are the only people who notice. Within a week, the child has forgotten the incident and Mario has died of a sudden heart attack. Would it be conducive to well-being for you to blame Mario? Your blame won't incentivise better behaviour from Mario in the future, nor will it help the child. It will make you significantly less happy if your otherwise positive memories of your friend are clouded by hostile feelings towards him. It will also make you less likely to say the things that need to be said to Mario's bereaved family in order to lessen their pain: that he was a good man, that you will remember him with nothing but fondness, and so on. Yet Mario did something that he should not have done – it hurt the child and failed to make him feel any better, it was not part of an optimific option for him (it came from his commitment to profit from the arms trade, which is, we can assume, not one that utilitarianism would endorse), and he had no excuse. So this is a case where self-effacing utilitarianism holds that someone did something they shouldn't have, without excuse, and you should not blame them for it.

Such a conclusion is perhaps defensible in cases like Mario's. It is also implied by utilitarianism of a non-self-effacing kind. Stranger is the idea that one should sometimes blame someone even though they have done as they should. Utilitarianism of all kinds infamously implies this. The usual case involves a sheriff, who finds that it would be optimific to convict an innocent person when acquitting them would provoke the mob into a destructive riot. A more quotidian one involves children. Imagine an older sister who climbs a tree to retrieve a kite for her younger brother. Her action was conducive to well-being but breached a rule set down by their father that neither child should climb trees. Upholding this rule is important and optimific, because the younger brother could not climb a tree safely and would not accept a rule that applied to him but not to his sister, who can. It

might be most conducive to well-being for the father to blame the sister for her actions, thereby strengthening the rule – even though what she did brought happiness.

Self-effacing utilitarianism is in a worse position than subjective utilitarianism, however, as it implies that one should sometimes blame people not only when they have done as they should, but also when blaming them *will not* be conducive to well-being. Very plausibly, for those commitments in the category of moral convictions, having a commitment involves having dispositions to praise and blame certain actions. For example, commitment to 'thou shalt not steal' requires not only refraining from theft, but also blaming those who steal, praising those who do not. For instance, if somebody who was generally disposed not to steal and felt guilty about doing so stole in order to feed their children in circumstances that gave them few alternatives, blaming them would probably not add to the sum of well-being. If this act of blame was a necessary part of a set of actions, involving having the commitment to 'thou shalt not steal', which was conducive to utility, then this case is like George's: self-effacing utilitarianism does not ask us to refrain from blaming, even though it is not conducive to well-being to do so. But the theft in this case might well have been conducive to well-being; it might even be what utilitarianism requires the thief to do. (In any case, it seems well excused.) Therefore, self-effacing utilitarianism seems to hold that an agent should be blamed for doing something that they should have done, or have an excuse for doing, even when blaming them has a negative effect on well-being.

Just like blaming, advising can, self-effacing utilitarianism implies, come apart from judgments about what should and should not be done. Utilitarianism of all kinds can hold that one should advise someone not to do something that they should do, when this would be conducive to well-being. Recall the case of the tree-climbing older sister. It would be conducive to well-being, plausibly, for her father to tell her not to climb the tree (thereby upholding the rule), even though well-being would be served by her doing so.

Self-effacing utilitarianism may further imply that such advice should be given, even when giving that advice would not be conducive to well-being. If I am committed to the commandment 'thou shalt not steal', then I will be disposed not only not to steal and to blame those who do, but also to advise

others not to. If the parent in the case above asked me for advice, I would respond that they should not resort to theft. But their taking this advice might lead to worse consequences for them and their children's well-being. And yet I would have to give that advice as part of my commitment to the commandment. Self-effacing utilitarianism therefore implies cases in which we should advise people not to do things that they should do, even when giving such advice is not conducive to well-being.

To recap: the judgment that one should X ought to have force, i.e. encourage certain responses of motivation, blame and advice. Self-effacing utilitarianism makes it hard to explain how the judgments it issues motivate us, unless we are systematically deceived or utilitarianism is indeed ushered from the scene. It also implies that there are cases in which one should blame people for doing as they should, and advise them not to do as they should – even when such blame and advice is not conducive to well-being. How significant these problems are is unclear. Some similar distinctions do not seem so problematic. For some people the judgment that one should not do something triggers a motivation to take violent retribution against those who do it. They seem to be wrong: the judgment that one should not X need not imply that X should be violently punished. But the judgment that one should X does seem to imply, at least weakly, that one should be motivated to X, and advise rather than blame X-ing. If the judgments of self-effacing utilitarianism do not do these things, then we might say that they are not really moral judgments, and utilitarianism, therefore, has ushered itself from the moral scene.

5.4. The belief problem

As a utilitarianism, self-effacing utilitarianism endorses the claim that one should choose the option that will result in the most well-being. However, it differs from subjective utilitarianism in recommending that one is *not* motivated by a concern to maximise well-being, and that one does not answer the question 'what should I do?' through utilitarian reasoning. We have already seen problems arising from these features in the last two sections. Here is a third: it seems difficult to resist ever being motivated by utilitarian considerations, or employing them in reasoning, if one believes utilitarianism. If one sincerely believes that one should do the thing most

conducive to well-being, then how could one not employ this principle in deciding what to do, at least in difficult situations where the answer does not present itself from any of one's other principles or commitments? And if one believes that one should do certain things, it is normal to thereby be motivated to do them – indeed, this is the phenomenon that self-effacing utilitarianism struggled to explain in the previous section, giving rise to the force problem.

Now, which beliefs one holds is also a choice between options that can yield different amounts of well-being, and therefore is up for utilitarian evaluation. If, as self-effacing utilitarianism claims, it is more conducive to utility for us not to be motivated by or reason with utilitarianism, and having a belief in utilitarianism will make such behaviour likely, this belief will likely not be recommended by utilitarianism. So self-effacing utilitarianism is in fact even more self-effacing than the last chapter made out. It is not just that it does not require us to employ utilitarianism, it also recommends that we disbelieve it. This is the belief problem.

Just as we might think that a moral theory should be such that we can be motivated or guided by it, we might think that it should be such that we can, according to the theory itself, believe it and its judgments. If so this result is fatal for self-effacing utilitarianism. However, Railton and Parfit deny that a moral theory must be believable by its own lights. Parfit defines 'self-effacingness' in a narrower way than I do, as susceptibility to the belief problem. But he does not see this as a significant problem. He writes:

> 'According to C, each of us should try to have one of the best possible sets of desires and dispositions, in Consequentialist terms. It might make the outcome better if we did not merely have these desires and dispositions, but had corresponding moral emotions and beliefs… If these claims are true, C would be self-effacing. It would tell us that we should try to believe, not itself, but some other theory… If a moral theory can be straightforwardly true, it is clear that, if it is self-effacing, this does not show that it cannot be true.' (1984, 41–43)

Railton writes:

> 'if maximising the good were in fact to require that consequentialist reasoning be wholly excluded, would his refute consequentialism?… On the contrary, it shows that objective consequentialism has the

> virtue of not blurring the distinction between the truth-conditions of an ethical theory and its acceptance-conditions in particular contexts, a distinction philosophers have generally recognised for theories concerning other subject matters.' (1984, 155)

Railton seems to be appealing, here, to the work of Bas van Fraassen (1980), in which the view known as 'constructive empiricism' is advanced. In van Fraassen's view, scientific theories should be construed literally, as making claims about the way the world – including the unobservable world – is. Theories are true if and only if the world is in fact the way they say it is. There are various ways that the unobservable world could be which would provide the same observable evidence to scientists, so theories could 'agree in empirical content and differ in truth-value' (1980, 36). Whether a theory is acceptable depends primarily on its empirical content. To 'accept' a theory, for van Fraassen, is to believe not that it is true but merely that it is 'empirically adequate' (i.e. consistent with all observable phenomena), and also to commit to using the theory in explanation and research (1980, 12–13). One should accept a theory if and only if it is empirically adequate and useful in explanation and research. Theories can be acceptable without being true, as neither empirical adequacy nor usefulness require correct description of the unobservable. So the acceptance-conditions and truth-conditions of scientific theories are distinct.

How does this idea help self-effacing utilitarianism? The task is to reach Parfit's conclusion, that the fact that a moral theory tells us not to believe it doesn't imply that it isn't true, from van Fraassen's distinction between truth-conditions and acceptance-conditions. Van Fraassen's notion of acceptance, with its reference to evidence, explanation and research, is more appropriate to scientific than moral theories; it also does not involve belief in the theory (just in its empirical adequacy). We might use a broader notion of acceptance that holds that to accept some theory is to be prepared to use it as a premiss in reasoning, and to end inquiry into whether it is true (Harman 1986, 47). On this broader notion of acceptance, one way of accepting a theory is to believe it, but a range of other attitudes, including van Fraassen's description of acceptance in science also amount to acceptance.

In the usual cases of constructive empiricism, truth-conditions and acceptance-conditions come apart such that a theory which could be false

may nevertheless be acceptable. To defend self-effacing utilitarianism we need the inverse claim, that a theory may be true even though it is unacceptable in some context. Is this possible? Under van Fraassen's account it is. A true scientific theory could never fail to be empirically adequate, since if a theory makes a claim that is inconsistent with our experience, it makes a false claim and therefore is not true. But a true theory could be less useful, in some context, for the pragmatics of science than a false one: it may be too complex for explanatory work, for example.

Could moral theories also be true but not acceptable in some context? This is the claim made by the self-effacing response for utilitarianism: it is true that we should choose the option most conducive to well-being, but we should also disbelieve utilitarianism. Railton's defence of this claim by gesturing at van Fraassen's distinction is not compelling. There is a significant disanalogy between a scientific theory that is true but not useful enough to be acceptable, and self-effacing utilitarianism. In the former case, what makes the theory unacceptable is that it is not useful, and that is a reason not to accept it *according to the norms of constructive empiricism*. In the latter, what makes the theory unacceptable is that accepting it would not be conducive to well-being, and that is a reason not to accept it *according to the theory itself*. In van Fraassen's constructive empiricism acceptance-conditions are not just independent of truth-conditions, they are also independent of the theory under consideration. Whether we should accept some theory is a matter of its empirical adequacy and usefulness. This normative epistemology provides the criteria for the acceptance of descriptive scientific theories. This is not structurally analogous to a normative theory providing criteria for the acceptance of normative theories including itself, and finding itself wanting. Establishing this disanalogy does not prove that the belief problem is a blow to self-effacing utilitarianism. But it does block the move that Railton seems to make, of appealing to van Fraassen's distinction between truth and acceptance in science as a demonstration that it is no problem at all. The belief problem is distinctive.

It is not, however, unique to self-effacing utilitarianism. Any theory that places value on effects will be by its own lights unacceptable in some context. Parfit employs an imaginary case involving Satan, who:

> 'perversely causes belief in [the true] theory to have bad effects in this theory's own terms… Suppose that the best moral theory is Utilitarianism… Satan ensures that, if people believe this theory, this is worse [in utilitarian terms] for everyone. Suppose next that the best moral theory is not Consequentialist, and that it tells each person never to deceive others, or coerce them, or treat them unjustly. Satan ensures that those who believe this theory are in fact, despite their contrary intentions, more deceitful, coercive, and unjust… Given Satan's interference, it would be better if we did not believe the best theory.' (1984, 43–44)

Just like self-effacing utilitarianism, the second theory here would face the belief problem. Like utilitarianism according to the self-effacing response, if it were true it would be unacceptable according to its own standards. Parfit says that such a theory could be non-consequentialist. However, it must still find value in the effects of what we do. The theory Parfit describes says that we should never in fact deceive, coerce or treat unjustly, and Satan could make it so that our actions had those effects. Another theory might say that we should never *aim* to deceive, coerce or treat unjustly. Such a theory would not place any value on the effects of what we do, only on our intentions. Satan could not make it such that if we believed this theory and tried to follow it, this would be bad in the theory's own terms: the only thing that is bad in this theory's terms is aiming to deceive, coerce or treat unjustly. So it is not only utilitarianism that can face the belief problem, but it is not all moral theories either.

Furthermore, the self-effacing response to Williams claims not only that utilitarianism could be self-effacing in fanciful circumstances like Parfit's, but that it *is* self-effacing in the actual world. Someone who believed that it is fatal for a moral theory to be such that it implies, in our world, that we should not believe it, could dismiss self-effacing utilitarianism on these grounds whilst endorsing some other theory that could, in Parfit's Satan case, be self-effacing.

So self-effacing utilitarianism cannot be defended from the acceptance problem by analogy with scientific theories as regarded by constructive empiricism; nor can it be defended by a tu quoque that claims it is in a no worse position than other moral theories. Having dismissed some reasons

for thinking that the belief problem is not a problem for self-effacing utilitarians, here is one reason to think that it is.

If the defenders of self-effacing utilitarianism believe what they defend, then they believe that utilitarianism is true. But they also think, given self-effacingness, that one should not believe utilitarianism. Therefore, by their own lights, they are doing something they should not be doing; furthermore, by advocating utilitarianism they encourage others to do something they should not be doing. This might not undermine the truth of utilitarianism, but it would be an unwelcome result for its adherents.

One way of putting this point is that if utilitarianism is self-effacing, it could not be the case that we ought to believe it. If it is false (unless by some coincidence the true moral theory is also self-effacing and recommends believing utilitarianism), then we have no reason to believe it. If it is true, then we ought, according to the self-effacing response, to reject it. This is one interpretation of Williams's memorable question of whether 'utilitarianism is unacceptable, or merely that no one ought to accept it.' In a later work, Parfit (2011, 2:619) makes a similar point about normative nihilism, the hypothesis that there are no normative truths. It could not be the case that we ought to believe such nihilism, because if it were true, there would be no true claims of the form 'we ought to believe X'. This, for Parfit, tells against normative nihilism, although not conclusively.

When talking about consequentialism in *Reasons and Persons*, however, Parfit defends the view that even if we ought to not to believe consequentialism by its own lights, consequentialism could be the best theory. He justifies this by distinguishing two questions:

> 'It is one question whether some theory is the one that we *ought morally* to try to believe. It is another question whether this is the theory that we *ought intellectually* or *in truth-seeking terms* to believe.' (1984, 43)

But this still presents us with a dilemma: if utilitarianism is self-effacing and true, we can either believe it and be intellectually correct and morally bad or disbelieve it and be intellectually incorrect and morally good. One of the two oughts must be transgressed.

Utilitarian*ism* may be true for all that has been said. But utilitar*ians* are people with choices to make. If self-effacingness gives them a reason to stop being utilitarians, this is an embarrassing result. Parfit's comment suggests that truth is not all that matters to making such choices. This is another area of agreement between self-effacing utilitarianism and van Fraassen's philosophy of science. But constructive empiricism only cares about intellectual oughts; the suggestion here is that choice of moral theory might depend partly on moral oughts. When deciding on which theory we ought 'all things considered' to accept, we might have to consider both. To put it another way: would a theory that we both intellectually and morally ought to believe not be preferable?

The utilitarian might reject this and claim that theory choice is a purely intellectual matter. But I do not think they should. Which theory one believes in has consequences for well-being, and it is in such consequences that utilitarians find value, no matter how they are produced (famously). It would seem arbitrary to ignore the value or disvalue created by the choice of moral theories when all other actions, dispositions and so on are subject to the utilitarian calculus. This is not to say that the utilitarian is committed to choosing moral theories without reference to which are true or favoured by intellectual oughts. But if moral oughts have any weight, the belief problem gives some reason not to endorse self-effacing utilitarianism.

5.5. Conclusion

Williams, reflecting on Parfit's version of the self-effacing response, did not deny what he took to be Parfit's conclusion, that their self-effacingness does 'not necessarily mean that such theories should be rejected.' But, he continued,

> 'this still leaves problems of who is to accept such theories, and in what spirit; and if it is not possible that any, or many, people should accept them, what the status of the theory is, and the purpose of the theorist in announcing it.' (1988, 192)

In this chapter I have described some of these problems, and of possible responses to them by proponents of self-effacing utilitarianism. I find the responses wanting. If a moral theory is not to provide useful guidance or

have motivational force for agents, then it seems to have abdicated one of its main responsibilities. And although the fact that we should not accept a theory does not imply that it is false, the fact that a theory says that we should not accept itself is a further problem; even if we grant that this is not fatal, the inevitable conclusion for an advocate of a self-effacing moral theory is that they are doing something they have some reason not to – which should be seen as a cost.

In the final chapter, I try to outline an account that avoids these problems, but also responds to the concerns underlying the Integrity Objection, whilst remaining utilitarian. If such an account can be made to work, it will provide the apologist for utilitarianism with a better response to Williams.

6. How should a utilitarian live?

6.1. The combination problem

Sidgwick wrote that

> 'if experience shows that the general happiness will be more satisfactorily attained if men frequently act from other motives than pure universal philanthropy, it is obvious that these other motives are to be preferred on Utilitarian principles.' (1962, 413)

In chapter 4, I gave reasons to think that the antecedent is true. Employing utilitarianism is often not conducive to well-being; for example in cases of time constraints, mental constraints and biases and games. In addition, an important source of well-being (according to Parfit) is acting from our own strong desires, rather than from the general concerns of utilitarianism, and if Raz's constitutive incommensurabilities exist, some goods which are conducive to well-being (such as friendship) require us not to think about our decisions in utilitarian terms. If Williams is right in asserting premiss 4, having commitments is conducive to well-being, and having commitments involves acting from the motivations characteristic of them, which is incompatible with acting on utilitarian motives.

Therefore, utilitarians should think that it is often best to guide one's decision not by employing utilitarianism, but in other ways. However, in chapter 5 I showed that utilitarianism should not usher itself entirely from the scene of practical decision-making. To do that would jeopardise its status as a moral theory worth endorsing. So a defender of utilitarianism must make some place in our deliberation for both utilitarianism and other, non-utilitarian principles, dispositions and motives. The place given to utilitarianism must solve the guidance, force and belief problems introduced in chapter 5. The place given to other motivations must incorporate the considerations raised in chapter 4, and respond to Williams's concerns about integrity. How can agents combine utilitarian and non-utilitarian deliberation in their lives?

6.2. The Harean account

Smart's answer is this: utilitarianism gives a criterion of rational choice – we should choose the option most conducive to general well-being. But such criteria only apply to conscious choices, where we deliberate using our rational faculties. Other decision-procedures may be useful in contexts where our rational faculties are not engaged – when we act out of habit. He writes: 'When we act in such an habitual fashion we do not of course deliberate or make a choice.' (1973, 42) Therefore, he goes on, the utilitarian may endorse non-utilitarian decision-procedures in such cases. Utilitarianism prescribes, given time and mental constraints, that we inculcate certain habits and train ourselves to act habitually in a range of situations. But when an agent 'has to think what to do, then there is a question of deliberation or choice, and it is precisely for such situations that the utilitarian criterion is intended.' (1973, 43) As our rational faculties are engaged, we come to see the rules involved in habitual action 'as mere rules of thumb, and will only use them as rough guides.' (1973, 42)

Smart's account seems to cover cases of emergency well, and it is 'when [the agent] has no time for considering probable consequences' (1973, 42) that he sees as the paradigmatic occasion for non-utilitarian decision-procedures. He can also give a utilitarian case for spontaneity (1973, 44–45). In such cases, we should act from instinct or general rules, and this is no problem for utilitarianism since we are not trying to make a rational choice. But it is not plausible that other cases in which non-utilitarian decision-procedures are appropriate are ones where we are not thinking, or trying to think, rationally. Smart himself discusses a case of a person 'trying to decide between two jobs, one of which is more highly paid than the other, though he has given an informal promise that he will take the lesser paid one.' (1973, 43) The risk, if this person employs utilitarianism to make his decision, is that the temptation of high pay will bias his deliberation. Sticking to the rule 'keep your promises' may be more likely to lead him to make the decision that is, in utilitarian terms, right. So he should stick to that rule. But following such rules is not really a matter of habit, and certainly doesn't mean that we do not 'deliberate or make a choice'. When people follow the rule 'keep your promises', they usually entertain the idea of such a rule consciously, and consider their options, and infer from the rule that they perform the promise-keeping action. This is not like

slamming the brakes on when a danger arises, or wearing a white shirt every day. There is a structured deliberation: 'What should I do, X or Y? I made a promise to do X. Well, I guess I should do X then.' It is not plausible that someone would take a job out of habit, without deliberation – it is an important choice, demanding rational engagement, and usually permitting the necessary time to think about it. These are the circumstances in which Smart recommends employing utilitarianism. But he produces a case in which it is nevertheless better not to.

Smart's habit-versus-rational-choice account is not able to handle the case of games discussed above, either. The claim is that there is an area of decision-making – the playing of games – in which utilitarian decision-making is inappropriate even if the agent is able to reach the correct utilitarian answer. We must suspend our utilitarianism whilst we play games, and aim to win, not to maximise well-being. And this could not be simply a matter of training ourselves to have some non-utilitarian habit. Playing games often involves paying close attention to the situation, weighing up alternative courses of action, making rational predictions and so on. It demands rational engagement and making real choices – but not through a utilitarian decision-procedure.

Furthermore, many of the non-utilitarian principles from which we act cannot be regarded as mere rules of thumb, if they are to serve their purpose with respect to utility. Take a commandment such as 'thou shalt not steal', the adoption of which may lead one to bring about more well-being than a biased attempt to evaluate each potential theft in the utilitarian manner. Someone who has adopted such a commandment will feel guilty when they break it, and will blame others who do so. And it is often good that we have such responses, because their unpleasantness motivates us to follow the commandments, which is good for well-being in general. That is not the case with mere heuristics: Parfit might have thought himself inefficient if he took too long deciding what to wear in the morning, but it would be odd for him to feel guilty. The point is that for some non-utilitarian but utilitarianism-endorsed motivations, the attachment we have to them is deeper and more moralised than Smart's 'rule of thumb' model suggests.

Smart's picture is highly at odds with Williams's notion of commitments too. Acting from a commitment is not typically a mere habit. If one is committed to Catholicism, attending Mass is not something one does

unthinkingly, like Parfit's putting on his white shirt in the morning or washing the dishes after dinner. For some people attending Mass is like this, but this is a mark of their waning commitment to the faith. The committed Catholic takes themselves to have good reasons to attend Mass. Those reasons spring not from what they can do for the general well-being, but from who they are. They may think that they could do no other, that they would be incapable of missing Mass. But this does not show that their attendance is a blind habit. As Williams says, a moral incapacity is the output, not the input, of deliberation (1992, 64–65). They cannot miss it because they (take themselves to) have such-and-such a reason to go. Nor can they regard their regular attendance at Mass as 'a rule of thumb', something that helps them do the right thing, but is merely a guideline, an instrument. It is a rule around which they have built their life.

A commitment to Catholicism is probably not something that utilitarianism would recommend. But it shares important features with other commitments which are more plausibly endorsed by utilitarianism, such as friendship. The point is that commitments cannot be regarded as habits or rules of thumb, but involve non-utilitarian decision-making (the following of rules, motivation stemming from one's identity). Smart's account of how decisions are made in a utilitarian life will not allow us to bring commitments into it.

Hare (1981) distinguishes two 'levels' of moral thinking: the critical and the intuitive. Moral thinking on the critical level determines what we should and should not do, deriving the answer from the workings of moral language and the empirical facts of the situation. According to Hare, by a metaethical manoeuvre which does not concern us here, the logic and linguistics of moral judgment and expressions imply that one ought, in any situation, to act so as to bring about the most satisfaction of preferences, impartially weighted, i.e. to follow the recommendations of utilitarianism. Since we cannot go back to those first principles in every decision we take (Hare's reasons for this have already been noted in chapter 3), we also need to think on an intuitive level. On this level, we employ general principles such as 'thou shalt not steal', deciding what to do by applying such principles to our situation (i.e. asking ourselves whether this would be a case of stealing). 'Having' a principle such as this, for Hare, means 'having the disposition to experience the feelings' associated with what Ross calls 'compunction' – the

obligation to do something, and of guilt or remorse having failed to do it (1981, 39). Which principles one has and how well one complies with them is crucial in determining one's moral character, whether one is (in a phrase Hare often uses) 'well brought up' or not.

Hare's picture therefore differs from Smart's in that it takes non-utilitarian decision-procedures seriously as moral thinking, not merely as unthinking habits or rules of thumb. They do engage our rational faculties; they do connect to moral feelings and practices such as blame; we can be attached to them as a matter of our identity and character. When does Hare think we should use our critical, rather than our intuitive, moral thinking? In the cases discussed in chapter 4, where it is more conducive to well-being to think in a non-utilitarian manner, we should clearly use our intuitive thinking, and if we were to consult critical thinking it would tell us to do so. But critical thinking is needed, according to Hare, to solve higher-order questions: which intuitive (or 'prima facie' (1981, 45)) principles should we employ? And how should we respond to conflicts between them? (1981, sec. 2.5) Critical thinking is also the more appropriate tool for considering extraordinary cases, which differ so far from the everyday that our intuitive principles – derived as they are from upbringing and experience – cannot be expected to handle them (1981, sec. 8.2).

Hare writes: 'For the selection of prima facie principles, and for the resolution of conflicts between them, critical thinking is necessary.' (1981, 45) There is some unclarity in how he suggests that critical (i.e. utilitarian) thinking selects intuitive principles. At one point he writes that

> 'critical thinking aims to select the best set of prima facie principles for us in intuitive thinking… The best set is that whose acceptance yields actions, dispositions, etc. most nearly approximating those which would be chosen if we were able to use critical thinking all the time.' (1981, 49–50)

Hare thinks that the reason we cannot use critical thinking all the time is primarily pragmatic – that we need general principles and moral learning to cope with decision-making, given our limited human minds. What utilitarianism is useful for, then, is selecting which principles would be best, given this 'depressing truth about reality' (Parfit 1984, 45). The answer in this passage seems to be that the best set of principles is the one that leads

to us living in the way that is closest, given these limits, to the 'archangel' (Hare 1981, 44), a being without such limits who employs utilitarianism in every decision it makes. One uses utilitarianism to work out how the archangel would act, and then works out from anthropological facts which principles would be best for driving humans towards those actions. But Hare follows the passage quoted above by saying that this claim 'can be given in terms of acceptance-utility' (1981, 50). Here one would work out which are principles such that their adoption would be most conducive to the general well-being. Choosing the principles that will most closely approximate the archangel and choosing the principles whose acceptance yields the greatest utility are not – as Hare seems to overlook – the same thing. Consider again the case of games. The archangel – using utilitarian critical thinking in every decision – would find it impossible to engage in competitive games, because as we saw in-game decisions should not be made by utilitarian calculation. This is not a matter of mental limits (archangels do not have any). If we were to choose the *prima facie* principles that made our actions as close as possible to those of the archangel, we would be unable to play games. Assuming this would be bad for general well-being, the principles with the highest acceptance-utility would differ.

The more thoroughgoing utilitarian response to this dilemma is to recommend choosing the principles with the highest acceptance-utility, even if this takes us further away from the archangel. To sacrifice well-being for the sake of being more angelic smacks of the self-indulgence that utilitarians condemn in virtue ethics. The charge in that case is that virtue theorists are more concerned with an agent's coming up to some standard than they are with the well-being of their fellow persons. Whether this is a compelling charge against the virtue theorist (Williams argued that it was not (1981)) I leave open. But a utilitarian should see its force, since their theory is founded on the importance of well-being. A utilitarian should approach the decision about which principles to adopt in the way that utilitarianism evaluates all decisions: by asking which option, of all those available, would be most conducive to the general well-being.

So my Harean utilitarianism recommends acting from non-utilitarian principles in the cases from 4.1, when doing so leads to the most well-being; it also recommends deliberating in a thoroughly utilitarian way when it

comes to selecting this principles. How does this picture differ from rule-utilitarianism? Rule-utilitarianism holds that we should act in accordance with the set of rules whose adoption would maximise well-being. Hare is not a rule-utilitarian. He writes:

> 'there is no harm in saying that the right or best way for us to live or act either in general or on a particular occasion is what the archangel would pronounce to be so if he addressed himself to the question.' (1981, 47)

This suggests that, against rule-utilitarianism, we should choose the options most conducive to the general well-being, no matter their conformity to any rule, since the archangel reasons purely by employing utilitarianism. Hare equivocates on whether the criterion of right action is act-utilitarian, as implied in this passage, or rule-utilitarian. Later in the book, he considers the case commonly put against utilitarians, that of the doctor who kills one person in order to use their organs to save five others. He writes: 'If we are to do the intuitive thinking, the matter is fairly simple. It *is* murder, and *would* therefore be wrong. [Hare assumes that 'murder is wrong' is a good prima facie principle.] A utilitarian does not have to dissent from this verdict on the intuitive level.' (1981, 132) This suggests that there is a sense of 'right' and 'wrong' appropriate to the intuitive level, such that it can be true that what the doctor did was wrong, even though it is what the archangel would do. This sense accords with rule-utilitarianism. However, in section 4.4, a criterion of right action can come apart from the question of what we should do.

In the case of actions that are wrong on the intuitive level but are conducive to well-being, what should the utilitarian do? Utilitarianism is committed – as I have defined it – to the judgment that one should choose the option most conducive to well-being. As we saw in chapter 4, when an action which is not optimific is part of an option that is, this implies that utilitarianism will sometimes recommend performing actions that are not conducive to well-being. But putting those cases aside, imagine that an agent truly is faced with two options consisting only of performing or not performing some action, which is conducive to well-being and is wrong according to some prima facie principle which utilitarianism recommends that we have. Utilitarianism must ask her to perform the action; rule-utilitarianism asks her not to.

This is also how a utilitarian agent must look at such choices. One cannot be a utilitarian and think that one ought to knowingly choose options that are not conducive to well-being. This is not simply a matter of being the purest, truest believer in utilitarianism. It is a matter of holding onto what is attractive about the doctrine in the first place. Utilitarianism draws on the pervasive desire to be happy to suggest that there is something off about not wanting the greatest total well-being. For me, the attraction of utilitarianism is its concern for something undoubtedly real and valuable – well-being – over the instruments of social control and inventions of moral philosophers (rules, laws, virtues, principles and so on). If one is minded to sacrifice well-being, knowingly, for the sake of conforming to a rule, then it is difficult to see why one would want to claim the label 'utilitarian' at all. As Scanlon puts it, 'philosophical utilitarianism' holds that 'the only fundamental moral facts are facts about individual well-being… it is the attractiveness of this doctrine which accounts for the widespread influence of utilitarian principles.' (1982, 108)

I also think that in many (perhaps most) cases of moral decision-making agents never get to a point where they recognise what it is that utilitarianism recommends. Well-being is notoriously difficult to measure, let alone predict, and all but the most straightforward decisions involve interpersonal comparisons and the consideration of remote effects and possibilities. Given that one does not know what utilitarianism recommends, one can be a utilitarian and follow some other decision-procedure; if that procedure is a prima facie moral principle (justified by utilitarianism) telling you that you should do something, then one has some moral duty to follow it. So in cases of conflict between utilitarian and intuitive thinking, one may be permitted and even required to follow the intuitive recommendation, when one does not know what the utilitarian recommendation is. When one does know the latter, retaining the essence of utilitarianism means following it.

Thus, in cases of emergency and bias, it is usually best to act without employing utilitarianism, because one doesn't know what the utilitarian recommendation is (because one is unable to figure it out). However, if somehow – perhaps by something like divine revelation – one did come to know which action would be most conducive to the general well-being, then if one were a utilitarian one would perform that action. The game-playing

cases are harder. Getting the most out of competitive games requires that one doesn't always act in the way most conducive to general well-being. But consider that in such cases, it is usually very difficult to *know* what utilitarianism would recommend. Returning to the penalty-taker, to know the utilitarian recommendation would necessitate making accurate predictions about the effects on the mental states of not only the players on the pitch, including herself, but also of large numbers of fans, on both sides, who are strangers to her. And note, in the cases where it is plausible that the agent does know what utilitarianism recommends, and this is at odds with the principles of games or avoiding extortion, it is not intuitive that the utilitarian option is the wrong one. In games, the aim of winning can be set aside or dialled down. Imagine a father playing chess with his son whom he knows will be very unhappy if he loses, so makes a deliberate mistake; a football team who realise early in the game that they are far superior to their opponents so play a little less hard to make the match more enjoyable for both sides; a bowler, in cricket, who refrains from bowling bouncers at a tail-end batsman's head to eliminate the risk of serious injury, though it might cost a few runs.

The picture of a utilitarian life that I am suggesting, then, goes like this. Utilitarianism is employed to select non-utilitarian decision-procedures, which are then employed in normal circumstances but overridden when one knows what the utilitarian recommendation is. Agents should adopt utilitarianism as a project and allow it to motivate their actions. Given that they are committed to some prima facie principles, and these should be overridden when one knows utilitarianism recommends it, they should also be committed to utilitarianism. It would be strange, difficult and alienating to override principles one is committed to for the sake of a theory one was not committed to. But one would not, in the ordinary run of things, employ utilitarianism to make one's decisions, and need not keep one's commitment to it in one's mind at all times. In these ways, utilitarianism is very much on the scene of practical decision-making, but waits in the wings rather than stealing the show, allowing space for non-utilitarian responses to cases of time constraints, human limitations and games.

My picture is Harean, in that it distinguishes two modes of moral thinking, giving utilitarian deliberation a higher-order role of choosing the procedures by which one tends to live. It resolves the unanswered question

in Hare about conflicts between two levels, holding that if one knows the utilitarian recommendation, one must act in accordance with it, even if that is at odds with some prima facie principle one has adopted. It goes beyond Hare in making the non-utilitarian 'intuitive' level consist in more than prima facie principles: there are all sorts of decision-procedures and motivations that could be appropriate, which do not look quite like general moral principles, and which utilitarianism can evaluate and select – for example, commitments. It also makes 'critical' level thinking a little less complex than Hare sometimes took it to be. In my picture, this thinking is characterised just by its employment of utilitarianism; for Hare, it can involve deep dives into moral semantics from which he derives utilitarianism. I say nothing about the derivation of utilitarianism.

6.3. The problems solved

My Harean utilitarianism solves the problems of guidance, force and belief that a wholly self-effacing utilitarianism could not.

The guidance problem was that if utilitarianism is a good moral theory, it must be able to help us answer the question 'what should I do?', and not only by supplying answers as if from a black box, but by showing the way to reach them. Self-effacing utilitarianism cannot do this without deception because it holds that we should not think in the manner in which utilitarianism reaches its answers. In my account, this is not the case. Sometimes – as in cases of the selection of prima facie principles – we should be straightforwardly guided by utilitarian concerns, selecting the principles that will result in the most well-being because they will result in the most well-being. More often, according to my account, we should reach our answers using those principles (and dispositions, desires, and so on) rather than utilitarianism: but if we were to ask why we should deliberate in this way, the theory would supply an answer – because doing so is most conducive to well-being. The agent who consciously abides by the commandment 'thou shalt not steal' is offered a justification, if she wants one, in utilitarian terms: the black box is transparent.

The problem of force is that the judgment that one should do something ought to provoke certain responses from the agent; specifically, one should be motivated to do what one should, advise others to do so, and blame those

who do not. For self-effacing utilitarianism, it is unclear why people should be motivated to do as they should, since whether one should do something is determined by its effects on well-being, but these effects should not motivate us. In my Harean account, agents should do what will maximise well-being, if and because they know it will. Furthermore, it is transparent to them (on reflection) that what we should do is determined by what will maximise well-being. So they will be motivated to do something if they know that they should, because they know that this means doing so will maximise well-being, and they are motivated to do so.

My account does not escape the problems of advice and blame, however, which affect all versions of utilitarianism. The problem is that there are cases in which advising someone to do what utilitarianism says they should, or blaming them for doing what utilitarianism says they shouldn't, would not maximise well-being (and cases in which advising someone to do what utilitarianism says they *shouldn't*, or blaming them for doing what utilitarianism says they *should*, *would* maximise well-being). Judgments about what we should do are not therefore straightforwardly connected with what we should advise or blame. However, this result is not as embarrassing for my account as it is for self-effacing utilitarianism, since my account has a different way (motivation) of giving force to its judgments.[16]

The belief problem is that self-effacing utilitarianism holds that if utilitarianism is true, we should not believe it. This doesn't undermine the truth of the theory, but it would be a cost to utilitarians, as they could not escape doing something they should not (because they either believe a false theory, or else the true theory says that they should believe otherwise than they do). According to my Harean account, we are permitted to believe utilitarianism. In fact, we probably should, since we ought to employ it in important deliberations such as the selection of prima facie principles, and if we disbelieved it then it would be difficult to justify doing so to ourselves. The account does hold that we should sometimes act from other principles, including ones that contradict utilitarianism. To allow this, we should not

[16] Hare also has a way of explaining blame (and connected feelings and practices such as remorse, praise and guilt): it relates to the intuitive level of moral thinking, so that we tend to blame people for breaking prima facie principles – indeed, it is part of having such a principle that one blames those who transgress it. I will not argue for this here.

keep our belief in utilitarianism in the front of our minds at all times. But that is true of most beliefs. I will respond to the suggestion that this is problematic in section 6.5.

6.4. Commitments revisited

Hare believed that his account accommodated Williams's concerns. Discussing the case of Jim, he writes:

> 'Professor Bernard Williams… thinks that he can score against the utilitarians by showing that in this far-fetched case they would have to prescribe the killing of one innocent man, the alternative being that he and nineteen others would die by another hand. We all have qualms about prescribing this – very naturally, because we have rightly been brought up to condemn the killing of innocent people, and also to condemn succumbing to blackmail threats of this sort, and good utilitarian reasons can be given to justify such an upbringing. But when we come to consider what actually ought to be done in this bizarre situation, even Williams seems at least to contemplate the possibility of its being right to shoot the innocent man to save the nineteen other innocent men… All he has shown is that we shall reach this conclusion with the greatest repugnance if we are "decent" people; yet there is nothing to stop the utilitarian agreeing with this.' (1981, 49)

Interpreted thus, the Integrity Objection merely demonstrates our strongly held negative feelings about breaking our moral commitments. Hare's two-level view can explain such feelings without giving up anything on the part of utilitarian critical thinking. Jim has these feelings because he is forced to breach a principle he has on the intuitive level of moral thinking. Those feelings are not irrational or immoral because having such principles helps him to act correctly (by utilitarian lights) most of the time. So if utilitarianism is the right theory to endorse on the critical level, Jim should shoot the innocent man, and he should feel bad about it (and Williams's readers should feel reluctant to endorse such action). This is, according to Hare, exactly the conclusion that Williams himself has, and so utilitarianism is in no way inconsistent with Williams's concerns.

But Williams's aim is not to give a counterexample, in which he 'enlist[s] the sympathies of [his] audience on [his] side by showing that the utilitarian is committed to views which nearly everybody finds counterintuitive.' (Hare 1981, 130) The examples of George and Jim are rather meant to focus our attention on a feature of moral life – commitments – which utilitarianism struggles to grasp. Here he explicitly rejects the characterisation Hare later made of his argument:

> 'The point here is not, as utilitarians may hasten to say, that if the project or attitude is that central to [Jim's] life, then to abandon it will be very disagreeable to him… on the contrary, once he is prepared to look at it like that, the argument in any serious case is over anyway. The point is that he is identified with his actions as flowing from projects and attitudes which in some cases he takes seriously at the deepest level.' (1973, 116)

So can my Harean account respond to Williams's objection? Recall my reconstruction of Williams's argument:

1. Having a project as a commitment
    a. is incompatible with regarding that project impartially or as dispensable;
    b. requires certain action robustly with respect to changes in rewards.
2. Utilitarianism requires us to
    a. regard all projects impartially and as dispensable;
    b. act in a way that is sensitive to rewards.
3. By 1 and 2, utilitarianism requires us not to have commitments.
4. Utilitarianism requires us to have commitments.
5. By 3 and 4, utilitarianism requires us both to have and not have commitments: therefore, utilitarianism is incoherent.

My Harean account solves the problems that beset wholly self-effacing utilitarianism because it is only partly self-effacing. Although it acknowledges that sometimes we should think in non-utilitarian ways, it recommends that we sometimes act as subjective utilitarians. As we saw in chapter 4, wholly self-effacing utilitarianism would answer Williams's Integrity Objection by denying premiss 2. Some partly self-effacing utilitarian accounts would deny premiss 2 as well (see section 5.1). What I

call a commitment self-effacing view holds that the utilitarian way of thinking could be employed in cases other than where commitments are involved, but still never asks us to regard our commitments impartially or as dispensable. My Harean utilitarianism is not wholly self-effacing or commitment self-effacing. One of the primary occasions on which utilitarian thinking should be employed, on this view, is in the selection and review of commitments. As we saw in chapter 2, utilitarian thinking about our commitments involves regarding them impartially and as dispensable. In reviewing our commitments, we would have to stand aside from them, ask whether having them is conducive to well-being, and be prepared to drop them if not. So 2a is true for my version of utilitarianism. In addition, it is part of my account that when an agent knows that a certain option is most conducive to well-being, they should choose that option. This makes our actions sensitive to rewards, as one could be offered a reward which would make an option predictably conducive to well-being, even if it involved acting at odds with one's commitments. Therefore, 2b is also true for my Harean utilitarianism.

If my Harean view cannot deny premiss 2, then to avoid Williams's charge of incoherence it must deny at least one other premiss. The most plausible way to do this is to deny the conjunction of 1 and 4: that is, to hold that if having a commitment means treating it as 1 describes, utilitarianism does not recommend that we have commitments.

Call a commitment that 1 is true of 'a Williamsian commitment'. The claim is that utilitarianism does not recommend that we have Williamsian commitments. This is true if and only if it is not conducive to well-being for us to adopt and maintain such commitments when we have the option to do so. As utilitarianism, including in my Harean form, only makes recommendations about options that are possible for us to realise or not, the claim is only about commitments that we, to some extent, choose. There may be commitments that it is impossible for us to abandon (perhaps a love for one's children), and some that it is impossible for us to adopt (perhaps a patriotic devotion to a foreign country). If falling in and out of love is involuntary, utilitarianism will never ask to do, or not do, either. Furthermore, on my Harean view, we are to regard commitments in the utilitarian manner that Williams is concerned about only when we review our commitments. If we have commitments that we cannot choose, there

would be little point in subjecting them to such a review. So whilst I deny premiss 1, my account need not differ from it with regard to involuntary commitments, which include some of our most deeply held ones.

Furthermore, it is thoroughly in tune with the Harean view to hold that much of the time, we should treat our (non-Williamsian) commitments in just the way Williams thinks is necessary. When we are not reviewing our commitments, we could regard them partially and as indispensable. A lot of our actions could flow from our commitments, bringing the happiness that Parfit describes as coming from acting on strong desires. The incompatibility between my Harean view and 1a comes only from the implication of my view that on some – perhaps very rare – occasions we should treat our commitments in a utilitarian manner.

On which occasions? If we reviewed our commitments every time we thought we had some reason to, they would arguably not deserve the name 'commitment' at all. For Cheshire Calhoun, commitments are distinguished by their 'high degree of resistance to reconsideration' (2009, 619). One might plan to go for a walk in the evening, but if it began to rain one would abandon that plan, and it would still be true that one really did have that plan earlier in the day. But if one commits to something, one goes through with it even when the going gets tough, when circumstances or one's desires change. Committed spouses would not set about reviewing their marriage after the first row, or a short period of boredom, or because they thought there was a small chance that it would not work out. In the first instance, they would take action to try to remove those problems. 'We measure depth of commitment', Calhoun writes, 'by what a person is prepared to do or to resist in order to see to it that the intention to engage persists.' (2009, 618)

But sometimes we do, and should, review our commitments. Because we are committed to them, we need specially strong prompts to do so. Pettit suggests that a moral theory provides a 'standby' (2015, 218–22). Our actions should usually be guided by our dispositions and commitments. But sometimes we receive evidence that we are not living as well as we could. Spouses row, not just once, but every day. A religious person comes across an argument against belief in God that they find impossible to refute, try as they might. People who know you start speaking ill of you. You become aware of great suffering in the world and realise that you are doing little to prevent it. At these moments – when 'alarm bells' ring (Pettit 2015, 220) –

you should subject your dispositions and commitments to review by the moral theory you endorse. For the Harean agent, these are the occasions on which utilitarian regard for commitments enters.

The incompatibility between my Harean view and 1b comes from the requirement that one choose the option that will maximise well-being if one knows what that is. For any action required by a commitment, there could be some reward so big that it would make an option that involved not performing that action more conducive to well-being. As we saw in 4.3, a utilitarianism that acknowledges the positive impact of retaining commitments on well-being and focuses on options rather than actions can make the threshold for such a reward very high, making a utilitarian agent's action less sensitive to rewards than might be expected. My Harean utilitarianism can make use of this, and it can go further. A Harean agent would not breach a commitment for the sake of a reward when that reward makes doing so conducive to well-being, considering the whole of each option. They would only breach a commitment for the sake of a reward when they *know* that this is the case. This is a subset of the occasions in which it *is* the case. Moreover, a bigger reward will usually be required to make an option such that one knows (or should know) that it is more conducive to well-being than simply to make it such that it is. So my Harean account makes the actions of utilitarian agents even less sensitive to rewards, and their actions with respect to their commitments correspondingly more robust.

But would it be more conducive to well-being for such actions to be completely robust against rewards – i.e. for us to have commitments that satisfy 1b? If it is, utilitarianism would recommend something that my Harean account makes impossible, and Harean utilitarianism would be, as Williams's charge goes, incoherent. I do not think it is. Imagine there are two ways in which George could be committed to opposing chemical warfare. In the first his commitment is Williamsian: he would not be party to chemical weapons research for the sake of any reward, however large. In the second his commitment is consistent with the Harean account: if a reward were offered that made it knowable that doing so would maximise well-being, he would accept such a job. There are two scenarios. In the first he is made such an offer; in the second he is offered the job but no reward. In the first scenario, a Williamsian commitment would lead him to refuse

the job, whilst a Harean commitment would have him accept – and the latter would, *ex hypothesi*, be more conducive to well-being. In the second scenario, he would refuse the job under either a Williamsian or Harean commitment. So it would seem that having the Harean rather than the Williamsian commitment is a one-way bet to more well-being.

This is a little too fast, of course. It may be that having the more robust Williamsian commitment leads to well-being gains that are independent of those associated with accepting or refusing the job. Having more robust commitments, it might be argued, gives better grounds for others to trust us, insulates us against the temptation to give up on things, and gives our lives more shape and meaning. Perhaps this is true, and it is an argument for having commitments with some robustness rather than none. However, since Harean commitments can be robust to a fairly high threshold, it is not obvious that they cannot provide these goods just as well as Williamsian commitments. The case that seems most favourable to the Williamsian position is friendship: it seems that to enjoy and provide a persisting and meaningful friendship, with all the well-being associated with it, one cannot be such that some reward would induce one to abandon one's friend. On my Harean account, one should do this when one knows that a reward would be more conducive to well-being than continuing the friendship. But ask yourself, what kind of reward would have to be on offer for one to know this? It is very difficult to conceive of a reward good enough, and highly unlikely that one would ever be offered. So one can have a Harean commitment to a friend, and be disposed to put the friendship ahead of any actual or even significantly possible reward. If one would abandon a friendship for a small reward, one which might be offered in a close possible world, then one can see how this would undermine the trust a friend could have in you, and thereby the well-being that the relationship can generate. But would the notion that there is some far-off possibility in which you would trade your friendship for a large reward affect the well-being your friendship brought about in the actual world? I find this unlikely. It would be like saying that a judge who would accept a billion-pound bribe to swing a case could not deliver justice in normal cases. The possibility is so small it does not seem to undermine robustness in a way that matters.

If the above argument is accepted, utilitarianism does not require us to have commitments that satisfy 1b. To fully deny the conjunction of premisses 1

and 4, I need to also argue that it does not require us to have commitments that satisfy 1a. On my Harean account, there are times – when we review our commitments – at which we should regard our commitments impartially and as dispensable. Would it be more conducive to well-being if we never did so?

I think not. Consider – to use one of Williams's examples of commitment – a committed Zionist. They were raised with this commitment, and do not tend to question it. Their commitment motivates them to take certain actions, such as regularly visiting Israel, and to rule out others, such as supporting anti-Zionist politicians. They proudly identify as a Zionist and publicly argue for their cause. The gradual stream of horrifying news stories emanating from Palestine and counter-arguments they struggle to find responses to eventually cause them to reflect on whether Zionism is a good idea after all, and whether they should remain committed to it. They decide to investigate this question in as unbiased a way as they can, regarding the fact that some people (including themselves) are committed Zionists as merely one input into the inquiry. As a good utilitarian, they decide that they will abandon their commitment if their investigation shows that retaining it is not conducive to well-being.

Now, by the end of this story during the inquiry our subject regards their commitment to Zionism impartially – as no more a consideration in favour of retaining the commitment than anyone else's – and plans to dispense with it if it is shown to undermine well-being. They are still – we can stipulate – committed to it, in the sense that whilst the inquiry is going on, they continue to let their Zionism guide their action and identify with it during the time they are not actively inquiring. If their inquiry turned out to vindicate Zionism, they would retain their commitment without us having to say that they ever lost it. A Williamsian commitment could not be treated like this, according to premiss 1. It seems to me that it would bring about more well-being, not less, if people thought about their commitments in this way more frequently, largely because they would drop commitments that undermined well-being.

It might be objected that whilst it is plausible that for political commitments such as Zionism it might be more conducive to well-being for us to hold them in a non-Williamsian manner, the same cannot go for all commitments. To resist the Integrity Objection, the utilitarian must show

that it is not conducive to well-being for us to have *any* Williamsian commitments. In section 3.2 I gave reasons to think that our having commitments is conducive to well-being. Are any of those reasons undermined if we hold our commitments in the way of the Zionist in the example above – prepared, in certain rare circumstances, to regard them from a utilitarian point of view?

A chief contribution of commitments to well-being is that they make possible certain relationships, such as friendships and romantic and familial ties. Would it be impossible to engage in such relationships if one ever regarded one's commitments to them in the manner of the Zionist described above? I think not. Take Railton's couple Juan and Linda (1984, 150–51):

> 'Juan… has always seemed a model husband. When a friend remarks on the extraordinary concerns he shows for his wife, Juan characteristically responds: "I love Linda. I even like her. So it means a lot to me to do things for her. After all we've been through, it's almost a part of me to do it."'

If we are to take him at his word, Juan has a commitment to Linda, and thanks to that has a relationship which improves the well-being of both of them. Railton goes on:

> 'But his friend knows that Juan is a principled individual, and asks Juan how his marriage fits into that larger scheme. After all, he asks, it's fine for Juan and his wife to have such a close relationship, but what about all the other, needier people Juan could help if he broadened his horizon still further?'

Here Juan's friend is pressing a broadly utilitarian objection to Juan's commitment: he could bring about more well-being if he cared less for Linda, and more for those in the most need. Juan replies:

> 'Look, it's a better world when people can have a relationship like ours – and nobody could if everyone were always asking themselves who's got the most need. It's not easy to make things work in this world, and one of the best things that happens to people is to have a close relationship like ours. You'd make things worse in a hurry if you broke up those relationships for the sake of a higher goal.

> Anyhow, I know you can't always put family first. The world isn't such a wonderful place that it's OK just to retreat into your own little circle. But still, you need that little circle. People get burned out, or lose touch, if they try to save the world by themselves. The ones who can stick with it and do a good job of making things better are usually ones who can make that fit into a life that does not make them miserable.'

In his response, Juan exhibits impartial regard for his commitment: he evaluates it without reference to it being his, or being to someone he loves, but as an example of a good kind of thing in the world, just like anyone else's similar relationship. He looks at it, in a sense, from without: asking what can be done to make the world better, and finding (conveniently) his relationship amongst such things, rather than reasoning from within his commitment, as he would when he takes care of Linda for no other reason than because it is her and he is her husband.

Railton next claims that Juan's

> 'motivational structure meets a counterfactual condition: while he ordinarily does not do what he does simply for the sake of doing what's right, he would seek to lead a different sort of life if he did not think it were morally defensible.'

Juan's commitment to Linda, therefore, whilst not usually driven by moral concerns, is subject to review by them. This suggests that if Juan had failed to come up with a satisfactory answer to his friend and vindicate his commitment, he would have left the relationship. In the moment just after the question was put, then, Juan began to regard his commitment not only impartially but as dispensable. Does this regard, which makes his commitment non-Williamsian, make their marriage impossible, or less conducive to well-being? It is unclear to me why it should do either. Juan's capacity to evaluate his relationship in a utilitarian manner does not seem incompatible with his first response to his friend, which affirmed its status as a commitment: he loves Linda; his actions towards her mean a lot to him; the relationship is part of his identity. Partial and impartial perspectives on his commitment co-exist. That Juan has a moral basis for his commitment to Linda aside from his love for her may well make their relationship more, rather than less, secure. There are moments when our love is temporarily

weak, and on such occasions being capable of motivation by impartial concerns can help us to avoid behaviour which would threaten our relationship, or the well-being of the parties to it (see Lillehammer 1997, 192). Lastly, there are cases in which abandoning close relationships is conducive to well-being, and so is what utilitarianism recommends. In these cases, it is good for well-being if agents identify this and are thereby motivated to change their lives. Someone who is prepared to evaluate their commitments as Juan does is more likely to do so, and this will be conducive to well-being. Williamsian commitments appear neither necessary to relationships, or useful to drawing the most well-being from them.

How about the other considerations from section 3.2? Since commitments, on my Harean account, can still be treated such that we need special reasons to abandon them and otherwise are motivated by them, they can still assist us with overcoming handicaps and pursuing political, artistic and scientific projects. A commitment held in a Harean way need not be one that we set aside when the going gets tough – and just as with the case of temporarily weakened love, having an independent moral basis for our commitments might in fact help us stick to long-term projects in the face of adversity. Moreover, a Harean utilitarian would drop or alter their commitments when circumstances made clear that they were not conducive to well-being: this would reduce the number of martyrs in defence of lost causes and artists with unfulfilled dreams that were the downside of commitments.

In sum, the Harean agent would not be able to have Williamsian commitments – those that met the conditions of premiss 1. However, they would still be able to commit to things, in the ordinary sense of the word. And there seems no reason to think that the form of commitment they could achieve would be less conducive to well-being – it may even have advantages. This resists the Integrity Objection. Utilitarianism would not insist on our having Williamsian commitments, and therefore the fact that an agent following utilitarianism in my Harean account could not have them does not render the account incoherent.

6.5. The integrity of the practical realm

My account recommends both a utilitarian and non-utilitarian attitude towards one's commitments, at different times. This may be thought

incoherent in another way: how can I have this double-minded attitude to my commitments, regarding them as related to me in two different and opposed ways? Williams makes this charge against Hare, when he writes:

> 'you cannot combine seeing the situation in that way, from the point of view of those dispositions, with seeing it from the archangel's way, in which all that is important is maximum preference satisfaction, and the dispositions themselves are merely a means towards that.' (1988, 190)

There is, perhaps, a kind of incoherence here. But it is not a problematic or unreasonable one. There are cases in which we do combine different ways of seeing the world, with one way being a means towards the goal of the other. Railton gives the case of a tennis player (1984, 144–45), who has the aim of winning as many matches as he can, but is underperforming. He is advised by an old pro that he will play better, and hence win more frequently, if he forgets about this aim whilst he played. He should play for the love of the sport, rather than to win. In taking this advice, is the tennis player being incoherent? On the one hand, he has an attitude towards winning, regarding it as being of paramount importance. On the other, he pushes this attitude from his mind whilst he plays, and tries to aim at something else. There is something strange about this way of thinking, compared to the simple coherent attitudes of philosophers' models. We may call it incoherence. Railton gives it a more dignified title: 'sophistication'. In any case, it serves its purpose just in case the player does win more frequently as a result. If so, it is a reasonable way of him going about his work, coherent or not.

Nagel gives further cases of sophisticated attitudes: 'The only way to run downstairs is not to try, you cannot make her love you by doing what you think will make her love you, you will not impress the interviewer unless you stop trying to impress him.' (1970, 132) Like the tennis player, Nagel's agents must have mixed attitudes; on the one hand aiming to do something but realising that in going about it they should forget their aims. The Harean agent displays a similar mindset. She takes maximising well-being to be of paramount importance. But she pushes this attitude away when she acts from commitments, in usual cases. She does this because she has taken advice, from her own utilitarian deliberations, on which commitments to adopt in order to maximise well-being, and her divided attitudes are a good

thing, coherent or not, insofar as they serve this purpose. When she hears Pettit's 'alarm bells' – evidence that her life is not serving this purpose – she switches back into critical, utilitarian attitudes. Similarly, if Railton's tennis player started losing matches, or Nagel's lovestruck agent realised that he was not endearing himself to the object of his love, they would re-evaluate their strategies.

The point of these remarks is not to argue that we must take a sophisticated approach to our commitments. The point is that we do take such an approach in other parts of our thinking, seemingly unproblematically, and therefore the burden of proof is on Williams's claim that it is impossible with respect to commitments. (Hare's own response to such a charge was to assert that he does his own thinking in such a manner: I hope to have broadened the sample size, at least.)

Perhaps Williams means to point to not some psychological impossibility in sophisticated attitudes, but to logical inconsistencies arising from them. Insofar as one has a commitment, one believes certain propositions; for example, one might believe that there is no amount of money that would compensate betraying a friend for. But insofar as one believes utilitarianism, one believes opposed propositions, such as that there *is* some amount of money that would compensate betraying a friend for (because everything is commensurable in the currency of well-being). This inconsistency is why the two attitudes cannot be combined.

Note that this is not the same as the Integrity Objection's charge of incoherence. The problem alleged there was that utilitarianism told us to do and not do the same thing, rendering its advice useless. Here, the problem is that the Harean agent believes that something is and isn't the case – it is quite clear what utilitarianism says is the case. Is it a problem for a moral theory that it recommends agents to be inconsistent in this way? If such inconsistency were impossible, then it would be. But we all have some inconsistent beliefs, so it is not. Nor is it true that it is always irrational to hold inconsistent beliefs. Imagine that one has good reason to believe each of three propositions. They are, however, jointly inconsistent. One can infer that at least one is false, but has no evidence as to which one. There does not seem to be sufficient reason to reject any one of the three propositions;

since one has good reason to believe each of them, it seems rationally permissible to go on believing all of them, despite the inconsistency.[17]

Although it is easy to hold inconsistent beliefs, it is harder to consciously affirm them both at the same time. Thankfully, the Harean agent would not do this. When they employ intuitive moral thinking, they would endorse the beliefs that are associated with their commitments – and may even be disposed to reject the contradictory belief implied by utilitarianism. It is only when prompted into critical thinking that the latter would be endorsed, and at this level the former would not be. Holding inconsistent beliefs at different times is not necessarily problematic. It is what happens when we learn. One could think of the transition to critical moral thinking as a learning process, where one comes to see that one's previous, intuitive belief was false. There seems to be nothing wrong about this. Of course, one might subsequently decide that it is nevertheless useful, and try to cause oneself to believe it again. Causing oneself to believe things can be difficult, but if a belief stems from a commitment of ours, that is, with a part of our identity, it should not be difficult to revert to it simply by refraining from questioning it again. It is not a good objection to my account, therefore, that agents following it would endorse inconsistent propositions: doing so is possible and not necessarily problematic. Perhaps a problem can be drawn out of the inconsistency of the Harean agent: but more work needs to be done than simply pointing the inconsistency out.

Wiggins writes:

> 'Is Hare's proposal simultaneously to inculcate ethical dispositions, to ingrain them into the formation of moral agents, and yet *subordinate* them to a way of thinking (the critical way) that is a stranger to them? If they are *not* to be subordinated, then intuitive thinking comes loose from critical thinking in a way that Hare cannot consistently contemplate. If these dispositions *are* to be subordinated, however, then there is trouble of another kind, namely the alienation of moral agents from that which critical thinking will butt in to require of them.' (2006, 189)

In my view (6.2), Hare's proposal is that we simultaneously develop dispositions, commitments and other motives that are non-utilitarian, and

[17] Thanks to Mark Kalderon for this example.

subordinate these to utilitarianism, in that it is for their utilitarian value that they are selected, and that they can be overridden for the sake of greater well-being. If they were not so subordinated, as Wiggins points out, we would fall into the problems I outlined for the self-effacing view in chapter 5 and solved with my Harean view in 6.3. But if they are subordinated, does this not lead to alienation, as Wiggins alleges, from our actions and commitments? This is only partly, and not fatally, the case. This is because utilitarian critical thinking ought not be 'a stranger to them', and neither would it 'butt in' all the time: agents should combine both utilitarian and non-utilitarian modes of deliberation. In this section I have tried to give reasons to think that such combination is possible.

The integrity that Williams was worried about, primarily, was the relationship between an agent, their projects, and their actions. In most cases my Harean utilitarianism preserves that integrity, though it sometimes asks us to step aside from our projects and consider them in an impartial light. (Even in this case we are doing so out of respect for one of our projects – utilitarianism itself.) But there is another kind of integrity that Williams assumed: the completeness of different parts of practical thought, such that our attitudes have to be consistent with one another. Sophistication, which is demanded by my utilitarian account, questions this assumption. One can flit between two different perspectives on one's commitments. The practical realm may be properly thought of, as Hare argued, as consisting of multiple levels.

6.6. Conclusion

Williams's Integrity Objection is both subtle and forceful. It does not simply say that utilitarianism demands too much, or the wrong thing, from us, but rather that it looks at the world in a way that threatens the very thing the theory was meant to prioritise: well-being. Williams is right to say that utilitarianism's detached and impartial view of things, which makes everything commensurable and therefore dispensable at the right price, is incompatible with commitments – in the following way: we could not be committed to anything if we took solely, or primarily, that viewpoint. If commitments are conducive to well-being, therefore, subjective

utilitarianism, which tells us to maximise well-being and to regard everything in a utilitarian manner, is incoherent.

However, subjective utilitarianism is not the only utilitarianism. Just because utilitarianism takes a certain view of the world does not mean that it requires agents to take do so. The self-effacing response claims that utilitarianism is not incoherent, because it is perfectly compatible with regarding our commitments in the way that Williams demands. This response, however, comes with major problems.

In this final chapter I have outlined a third way between subjective and (wholly) self-effacing utilitarianism. This view asks that we sometimes regard our commitments from the impartial utilitarian point of view, and therefore is incompatible with commitments as Williams describes them. However, I have given reasons to think that the commitments that are conducive to well-being can be had in a different way, consistent with my Harean account and without undermining their contribution to well-being. The distinctive feature of this account is that it asks us to hold different sorts of attitudes and deliberate in different kinds of ways at different times. Though this might be thought double-minded, or inconsistent, it is how we often conduct our practical thinking – and might be the best way to do it, if we want our lives and those of others to be as happy as they can be.

7. Bibliography

Calhoun, Cheshire. 2009. 'What Good Is Commitment?' *Ethics* 119 (4): 613–41..

Darwall, Stephen. 2010. 'Authority and Reasons: Exclusionary and Second-personal'. *Ethics* 120 (2): 257–78.

Foot, Philippa. 1972. 'Morality as a System of Hypothetical Imperatives'. *The Philosophical Review* 81 (3): 305–16.

Goodin, Robert E. 1995. *Utilitarianism as a Public Philosophy*. Cambridge University Press.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press.

Harman, Gilbert. 1986. *Change in View*. MIT Press.

Lillehammer, Hallvard. 1997. 'Smith on Moral Fetishism'. *Analysis* 57 (3): 187–195.

MacFarquhar, Larissa. 2011. 'How To Be Good'. *The New Yorker*, 29 August 2011.

Mill, John Stuart. 1971. 'Bentham'. In *Mill on Bentham and Coleridge*, 39–98. London: Chatto & Windus.

———. 2008. 'Utilitarianism'. In *On Liberty and Other Essays*. Oxford: Oxford University Press.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Oxford Clarendon Press.

Norcross, Alastair. 2006. 'Reasons Without Demands: Rethinking Rightness'. In *Contemporary Debates in Moral Theory*, edited by James Lawrence Dreier, 6–38. Blackwell.

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

———. 2011. *On What Matters*. Edited by Samuel Scheffler. Vol. 2. 3 vols. Oxford: Oxford University Press.

———. 2016. 'Can We Avoid the Repugnant Conclusion?' *Theoria* 82 (2): 110–27.

Pettit, Philip. 1993. 'Consequentialism'. In *A Companion to Ethics*, edited by Peter Singer. Oxford: Blackwell.

———. 2015. *The Robust Demands of the Good: Ethics with Attachment, Virtue, and Respect*. Oxford University Press.

Railton, Peter. 1984. 'Alienation, Consequentialism, and the Demands of Morality'. *Philosophy & Public Affairs* 13 (2): 134–71.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford University Press.

Scanlon, T. M. 1982. ‘Contractualism and Utilitarianism’. In *Utilitarianism and Beyond*, edited by Amartya Kumar Sen and Bernard Williams, 103–128. Cambridge University Press.

———. 1998. *What We Owe To Each Other*. Cambridge, Mass. ; London: Belknap Press of Harvard University Press.

Scheffler, Samuel. 1988. ‘Introduction’. In *Consequentialism and Its Critics*. Oxford: Oxford University Press.

Sidgwick, Henry. 1962. *Methods of Ethics*. London: Macmillan.

Singer, Peter. 1972. ‘Famine, Affluence, and Morality’. *Philosophy and Public Affairs* 1 (3): 229–243.

Smart, J. J. C. 1973. ‘An Outline of a System of Utilitarian Ethics’. In *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford University Press.

Wallace, R. Jay. 2017. ‘Trust, Anger, Resentment: On Blame and the Economy of Disesteem’. presented at the The Practical, the Political and the Ethical (Seminar Series), Institute of Philosophy, Senate House, University of London, May 23.

Wiggins, David. 2006. *Ethics: Twelve Lectures on the Philosophy of Morality*. Harvard University Press.

Williams, Bernard. 1973. ‘A Critique of Utilitarianism’. In *Utilitarianism: For and Against*, 77–150. Cambridge: Cambridge University Press.

———. 1981. ‘Utilitarianism and Moral Self-Indulgence’. Moral Luck: Philosophical Papers 1973–1980. December 1981.

———. 1988. ‘The Structure of Hare’s Theory’. In *Hare and Critics: Essays on Moral Thinking*, edited by Douglas Seanor and N. Fotion, 185–98. Oxford: Clarendon Press.

———. 1992. ‘Moral Incapacity’. *Proceedings of the Aristotelian Society* 93 (n/a): 59–70.