REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree MPhil Year 2005 Name of Author ECZEIN, N

COPYRIGHT

This is a thesis accepted for a Higher Degree of the University of London. It is an unpublished typescript and the copyright is held by the author. All persons consulting the thesis must read and abide by the Copyright Declaration below.

COPYRIGHT DECLARATION

I recognise that the copyright of the above-described thesis rests with the author and that no quotation from it or information derived from it may be published without the prior written consent of the author.

LOANS

Theses may not be lent to individuals, but the Senate House Library may lend a copy to approved libraries within the United Kingdom, for consultation solely on the premises of those libraries. Application should be made to: Inter-Library Loans, Senate House Library, Senate House, Malet Street, London WC1E 7HU.

REPRODUCTION

University of London theses may not be reproduced without explicit written permission from the Senate House Library. Enquiries should be addressed to the Theses Section of the Library. Regulations concerning reproduction vary according to the date of acceptance of the thesis and are listed below as guidelines.

A. Before 1962. Permission granted only upon the prior written consent of the author. (The Senate House Library will provide addresses where possible).

B. 1962 - 1974. In many cases the author has agreed to permit copying upon completion of a Copyright Declaration.

C. 1975 - 1988. Most theses may be copied upon completion of a Copyright Declaration.

D. 1989 onwards. Most theses may be copied.

This thesis comes within category D.

☑ This copy has been deposited in the Library of UCL

☐ This copy has been deposited in the Senate House Library, Senate House, Malet Street, London WC1E 7HU.

Freedom, OF THE WILL Responsibility & Moral Obligation

Nadine Elzein

University College London

MPhil in Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I hereby declare that all of the work within this document has been completed by myself, and is entirely my own work.

Signed:____ _____ Date: 1 - 4 - 05

Nadine Elzein

Abstract

This thesis addresses the classic problem of freedom and responsibility, focusing on the way that certain issues within metaethics might be seen to have a bearing on this issue.

Various theorists have appealed to the concept of reason in order to explain responsibility. I will begin by discussing reason-based views of responsibility, specifically supporting Susan Wolf's view against certain criticisms that have been levelled against it.

I will argue that in order to be held responsible for an act, the agent must bear the right kind of relation to the reasons for acting that she is subject to, and that such reasons must be considered "objective" in the sense of being independent of the individual's subjective aims and desires. I will defend the view that there can be reasons that are objective in this sense against Bernard Williams's argument that all true reasons claims must depend on subjective conditions.

I will argue, in particular, that *moral obligations* cannot be plausibly explained in the same way that Williams explains other reasons claims. This allows us to adopt the above explanation of responsibility at least for moral reasons. This means we can still account for the most important and interesting cases: we can still account for *moral* responsibility.

I will go on to argue for an alternative explanation of moral duties, influenced by Mill and Strawson. This involves explaining moral obligations, and the kind of normative pressure we associate with them, in terms of the justification we might have for adopting certain attitudes towards an agent in response to their acts. Such justification relates both to the objectivity of moral reasons, and to facts about the quality of the agent's will towards others.

I will build on this view of moral obligations in developing a positive account of the conditions of responsibility.

Contents

Introduction

The purpose of this discussion will be to examine the problem of free will and determinism, with the aim of developing plausible conditions for responsibility that reflect a plausible explanation of moral obligations, and of the kind of normative pressure we associate with them.

I will begin by considering various approaches to the problem of freedom and responsibility. I will defend the view that there is a fundamental link between the idea that there are reasons that are "objective" in the relevant sense, and the conditions under which we might consider an agent to be responsible. Ultimately, I will argue that responsibility requires that the agent bear the right kind of relation to such reasons.

With this aim in mind, I will go on to discuss what it might mean for reasons to be objective in the relevant sense, and to consider the kind of arguments that may make us sceptical of the claim that reasons could have the kind of objectivity required to support such a conception of responsibility. I will try to refute the view that all reasons claims have subjective conditions, arguing that we have good grounds to think that at least some reasons claims, in particular, *moral obligations*, cannot be plausibly explained in this way.

I will go on to argue in favour of an alternative explanation of moral obligations, influenced by Mill and Strawson. This account emphasises the importance of considerations about the quality of the agent's will towards others, and the way this relates to the kind of justification we might have for adopting certain attitudes towards an agent. This aspect of the account is also closely related to the idea that a system of moral norms can be considered to be objectively valid. I will build on this view of moral obligations in developing an account of the conditions of responsibility.

This discussion will be divided into four chapters. Whereas the first two chapters will discuss responsibility and reasons quite broadly, and the debate surrounding these issues, the final two chapters will be focused primarily on developing and defending positive accounts of these phenomena.

In chapter one, I will discuss the basic problem of freedom and determinism. I will focus on two prominent reason-based views of responsibility: those put forward by Susan Wolf and John Martin Fischer. I will look at Susan Hurley's attempt to refute Wolf's view, and will try to show that Hurley's argument fails, and that we have better reason to support an account of responsibility closer to Wolf's. Wolf associates responsibility with the capacity to respond to objective reasons.

In chapter two, I will defend the idea that there are objective reasons of the kind needed to support Wolf's account. I will focus, primarily on Williams's argument against external reasons claims, since the kind of objectivity required to support Wolf's account involves a commitment to reasons that are external as Williams defines it. I will argue that even if we were to find Williams's account plausible in explaining some reasons claims, it breaks down when we try to account for moral reasons. Moral reasons are the kind that matter most, since the really interesting questions of responsibility are those concerning *moral* responsibility.

The third chapter will support an alternative account of moral reasons, influenced by Strawson and Mill. On this account, the *attitudes* we are justified in adopting in response to facts about the quality of an agent's will towards others are fundamental when it comes to understanding moral duties and explaining the normative pressure associated with them.

The final chapter will draw on these considerations in order to develop conditions of responsibility that not only take into account Wolf's insights, but also relate in a relevant way to features that are fundamental in explaining moral obligations. I will also discuss the relation between these conditions and those looked

at in the first chapter, and will reconsider the question of whether we should consider responsibility to be compatible with determinism.

1 Reason, Determinism & Moral Responsibility

In this Chapter, I will discuss the traditional problem of free will and determinism. That is, I want to consider (roughly) the thesis that all events, including our own actions, are determined by prior causes, and whether this thesis should be taken to pose some threat to the concept of free will, and in turn, to the concept of moral responsibility. I want to examine some of the ways in which theorists have tried to resolve this problem. I will focus on the kind of solution that involves appealing to the concept of *reason* to help ground the notion of responsibility in light of the threat we might take determinism to pose. I aim to defend one of these reason-based approaches against certain lines of argument that have been put forward against it.

I will begin by discussing what we mean by the thesis of determinism, and exactly why this might be seen as a threat to the concept of freedom, and of responsibility. I will then look at various responses to this problem, starting with Harry Frankfurt's argument that determinism should not be taken to threaten responsibility in the way we tend to suppose. I will then go on to discuss two prominent reason-based views of responsibility – those put forward by Susan Wolf, and by John Martin Fischer.

Ultimately, I want to defend an account along the lines of that offered by Wolf. Susan Hurley draws on Frankfurt's point in order to argue that we should reject Wolf's account in favour of Fischer's. I hope to show that her argument fails. This is because it rests on an inference that commits a particular fallacy – one outlined by Robert K. Shope, which he calls "the conditional fallacy".

1.1 Freedom of the Will

1.1.1 The Threat to Free Will

The phenomenon most notably taken to threaten free will is causal determinism; the idea that the laws of physics and the past are enough to provide a full explanation of why any event happened, including events such as our own decisions. I will use the word "determinism" to refer to the following somewhat more precise thesis:

Determinism:

For any given time, a complete statement of the facts about that time, together with a complete statement of the laws of nature, entails every truth as to what happens after that time.

For the purposes of this discussion, I will not question the truth of this thesis, but will work on the assumption that it's true in order to examine the implications this has for the notion of responsibility[1].

This thesis entails that our own actions and decisions are determined by factors entirely outside of our own control. They are already implied by events prior to our birth. If our actions were already determined prior to our birth, we might think this implies that we cannot make free decisions in the way we suppose we can. This, in

---

[1] It's worth noting that even if we do *not* think that such a strong statement is true, we might take a slightly weaker thesis to threaten freedom and responsibility in much the same way. Some theorists argue that the laws of nature are not always entirely deterministic in this way, but might be, to some extent, probabilistic. Thomas Scanlon points out that even this does not necessarily help us to resolve the threat, since our actions would still be the result of causal influences outside of our control; it's just that they will affect us in a probabilistic rather than a deterministic way. ("Responsibility" in *What we Owe to Each Other* (Harvard University Press: 1998) p. 250). To simplify matters, I will work with the straightforward determinism case in this discussion, focusing purely on its relevance to the issue of moral accountability, and not on its truth or falsity.

turn, may lead us to conclude that we cannot justifiably be held responsible for our decisions. If the past and the laws of nature are enough to determine our actions, then given we cannot change either of these things, it seems it seems to follow that we can never do anything different to what we do in fact do.

The truth of determinism has been considered to be a threat to responsibility because it seems to imply that we are not able to do otherwise. It has generally been supposed that an agent can only be justifiably held responsible for her act on the condition that she is capable of doing otherwise, so if determinism is incompatible with this condition being met, it's also incompatible with responsibility.

For this reason, the debate about whether free will is compatible with determinism has generally focused on whether the ability to do otherwise is compatible with determinism. Traditionally, compatibilists have argued that it is, and incompatibilists, that it is not.

Compatibilists have, in the past, adopted a strategy known as the conditional analysis in order to argue that the ability to do otherwise is compatible with determinism. This involves a particular analysis of the following the statement:

(a) He could have done otherwise

It is argued that this should be analysed as being true on the condition that the following statement is true:

(b) He would have done otherwise, if he had chosen to do otherwise.

If this analysis worked, then the ability to do otherwise would be compatible with determinism, because statement (b) is compatible with determinism. But the analysis fails. As Hurley notes, the ability to do otherwise requires more than just acting on a disposition that would have led you to act differently under conditions that do not

obtain. It requires *the outright possibility of an alternate sequence of events holding all else constant.*

The conditional analysis fails because it does not distinguish features of the *actual* sequence from features of the *alternate* sequence. In assessing whether responsibility is compatible with determinism, we need to draw a distinction between actual sequence requirements for responsibility and alternate sequence ones. We need to know whether it's features of the actual sequence of events leading to the agent's act that count, or rather that we need the outright possibility of an alternate sequence of events with certain features.

An example of an actual sequence requirement is the regression principle, which states that in order to be responsible for an action, the agent must be responsible for its cause. An example of an alternate sequence requirement is the principle that an agent can only be held responsible for her action on the condition that she could have done otherwise.

Conditional analyses involve only actual sequence principles, and these can never be used to ground an ability to do otherwise. But even if the ability to do otherwise cannot be grounded in any actual sequence requirement, we might still think that responsibility *can* be. We can consistently maintain that some actual sequence principle provides a basis for responsibility, so long as we are willing to question the traditional assumption that ability to do otherwise is a condition of responsibility. Harry Frankfurt devised a case that suggests this traditional assumption may indeed be mistaken.

1.1.2 Frankfurt's Case

Frankfurt rejects the claim that alternate possibilities are necessary for responsibility. There are many cases where the agent cannot do otherwise *and* is not

responsible for their act, but this doesn't mean that the agent is not responsible *because* they cannot do otherwise. It may be that some factor, such as the agent's being coerced, would entail both of these things, but that there is no entailment between them. Normally it is the very same thing that stops the agent doing otherwise that also makes him perform his actual act. Frankfurt devises a case that involves no such presumption.

Black wants Jones to perform a certain action. He has a special device that will manipulate Jones's nervous system and brain processes, making him perform the action in question. But Black is excellent at judging people's intentions, and he will only bother doing this if he judges that Jones is not going to perform the act of his own accord. As it goes, Jones does perform the act of his own accord, so there is no need for Black to intervene.

In this case, it seems Jones is responsible because he acted entirely for his own reasons. It seems irrelevant that he could not have done otherwise. Hurley puts this down to the "irrelevant alternative intuition".

If we accept Frankfurt's argument, this opens up scope for accounts of responsibility that do not require any alternate sequence of events, but instead look only at features of the actual sequence of events leading to the agent's act. In fact, it might be taken to not merely make such a strategy feasible, but to add a substantial amount of support to it. If alternate sequences which play no role in determining the agent's actual actions are irrelevant in determining whether or not the agent can be held responsible, this means that principles which require the possibility of alternate sequences are equally irrelevant, and therefore only accounts that rest on actual-sequence requirements will do.

Hurley uses precisely this line of argument, based on Frankfurt's irrelevant alternative intuition, in order to support Fischer's reason-based view of responsibility,

which rests only on actual sequence requirements, and to reject Wolf's reason-based view, which in some cases involves alternate sequence requirements.

In the following section, I will examine both of these reason-based accounts. It is with reference to the precise conditions developed by Wolf and Fischer that Hurley is able to use Frankfurt's intuition to develop her argument, so we will need to look carefully at the conditions each of them offer. I will begin by looking at Wolf's view, and at the considerations she offers in support of it.

1.2 Reason Based Accounts of Responsibility

1.2.1 Susan Wolf's Account

Wolf compares three different models of responsibility: the Autonomy View, the Real Self View, and her own version of the Reason View, ultimately arguing for the Reason View. She analyses these as follows:

The Autonomy View

The Autonomy view is committed both to a regression principle and a could-have-done-otherwise principle. It requires radical freedom; that the agent has *ultimate control.* The agent's decisions must not be causally determined, but at the same time must not be random and uncaused. The self must be able to endlessly account for itself and its behaviour.

The Real Self View:

The Real Self view, in contrast, does not require that responsibility is regressive in this way – the regression stops with the agent's system of values, or real self. So long as this is the source of the agent's act, the agent is responsible. It does not matter

where the Real Self comes from. Wolf rejects this because an agent may not be responsible for their system of subjective values. They may result from a bad upbringing. She argues that this confuses mere causal responsibility with deep moral responsibility.

The Reason View:

It is the ability to act in accordance with objective reasons (as opposed to your own subjective values) that is required for responsibility. This is the view favoured by Wolf.

According to the autonomy view, an agent who does the right thing for the right reason will not be praiseworthy unless she could do otherwise. Wolf argues that this is irrelevant. It's just the agent's ability to act in accordance with reason that matters, not whether she is able to act irrationally instead. Wolf argues that in cases of praise the agent's being able to do otherwise does not add to his status as a responsible agent. In fact, it might even detract from it.

On Wolf's account, it's a mistake to think that an absence of determinism is always required for responsibility. She looks at precisely what it would mean for an agent to be able to do otherwise in cases of moral praise. Either they would have to be able to act in ways that contradict their own values – to act in *spite* of their values, or they would have to be able to be able to pick and choose their values. This would mean that those values could not possibly be based on the way things are, or on the agent's capacity to discern the actual value of things, but would instead be based on the agent's random whims.

Wolf argues that if an agent were able to act in ways that contradict their own values, this would be insane. E.g. a mother who could watch her son dying without helping him despite the fact she could, and that she loves her son and wants him to live.

Likewise, it would not help if someone were able to be in control of their values to the extent that they could drop them at a whim. E.g. a man who loves and cares about his wife, but whose love for her is such that he could just choose to stop caring about her at any moment. To say the least, this would not add any praiseworthiness to the acts he performs on the basis of his love for her. Arguably, this would not even count as genuine love at all. A person who could choose either to neglect or ignore their values might not even be considered a moral agent.

Wolf argues that the main point is that an agent's actions must be determined in the right way in order for the agent to be praiseworthy. It's not that they shouldn't be determined at all. An agent must be doing the right thing to some extent *because* it is the right thing to do. This does not require that the agent could do otherwise.

A bad upbringing does not stop a person being responsible in virtue of the fact it means their actions are determined; it does so in virtue of the fact that it means their actions are *not* determined by what reasons there actually *are*. Wolf reconstructs the conditional analysis as value-laden. An agent is responsible on the following condition:

> He could have done otherwise, had there been good and sufficient reason to do otherwise

In the case of praise, this condition is counterfactual, whereas in the case of blame it is not – the agent acts as they do *despite* the existence of better reasons not to. At any rate, the key factor is not our ability to autonomously govern ourselves, but our ability to be determined by the true and the good, as opposed to being determined by a random and misguided set of values.

Wolf's account is asymmetric between praise and blame. If we do the right thing for the right reason, this *a fortiori* implies that we were able to, and so the condition is automatically met. No outright ability to do otherwise is required. If we

do not do what reason requires, however, we cannot be held responsible unless we are *capable* of responding to reason. This *does* require the outright possibility of an alternate sequence of events.

It is this dependence on alternate sequence-principles for cases of blame (in the face of Frankfurt-style cases) that Hurley takes issue with, and that leads her to favour Fischer's account instead. Before going on to discuss Hurley's argument, we need to look at Fischer's model and at the way in which it differs from Wolf's.

1.2.2 John Martin Fischer's Account

Fischer's reason-based approach is somewhat different to Wolf's. He distinguishes between different *mechanisms* on which an agent might act. An agent is responsible so long as in the actual sequence, the agent is led to her act on the basis of a mechanism that is reasons-responsive. No possibility of an alternate sequence is required.

Jones is responsible because he acted on a reasons-responsive mechanism. Had Black manipulated his brain, he would not have been acting on a reasons-responsive mechanism, so he would not have been responsible. But as it goes, Black did not *actually* manipulate Jones's brain, so such a sequence of events is irrelevant to Jones's responsibility.

Fischer distinguishes between strong and weak reasons-responsiveness. Following Hurley, I will use the terms "tight" and "loose" reasons-responsiveness. A tightly reasons-responsive mechanism *tracks* reason in such a way that the same mechanism will always lead the agent to do what there is optimal reason to do. A mechanism is loosely reasons-responsive so long as there is some possible world (which need not be close to the actual world) in which there is reason to do otherwise, and the same mechanism operates, and leads the agent to act on that reason.

Weak-willed agents often do not do the right thing, but would do if there were slightly stronger reasons. Such agents will be responsible if all we require is loose responsiveness, but not if we require tight responsiveness. Fischer argues that tight reasons-responsiveness suffices for responsibility, but is not a necessary condition. An agent will still be held responsible despite some degree of weakness of the will.

Fischer's account of responsibility in most respects parallels Robert Nozick's truth-tracking account of knowledge. Nozick's account states that further to having true justified belief that *p*, an agent will only know that *p* if:

(1) the agent would *not* believe that *p* if *p* were not true (taking only nearby possibilities where *p* is false into account), and

(2) under various nearby conditions in which *p* were true, the agent *would* believe that *p*.

In other words, an agent has knowledge that *p* only if he is able to discriminate the conditions that would obtain if *p* were true from those that would obtain if *p* were false.

Stated in this form, where it's the capacity of the agent himself that counts, we run into problems. Nozick gives the following example:

> A grandmother sees her grandson is well when he comes to visit; but if he were sick or dead, others would tell her he was well to spare her upset. Yet this does not mean she doesn't know he is well (or at least ambulatory) when she sees him[2].

This example is an epistemological analogue to the Frankfurt case for responsibility. Nozick notes that this shows that making the conditions for knowledge

---

[2] Nozick, Robert. (1981) "Knowledge and Scepticism" in *Epistemology: An Anthology*, ed. by Ernest Sosa and Jaegwon Kim (Blackwell: 2000), p. 82.

agent-centred won't do. What's important is that we analyse the tracking capacity of the *mechanism* the agent bases her belief on. In the actual sequence the agent's belief is based on a reliable (truth-sensitive) mechanism, so the agent has knowledge. Had her son died, she would not have based her belief on a truth-sensitive mechanism, but since this alternate possibility plays no role in the actual sequence of events leading to the agent's belief, it's irrelevant.

For this reason we need a principle for knowledge that's both actual-sequence and mechanism-centred. Fischer argues, for analogous reasons, that we need an actual sequence, mechanism-centred principle for responsibility.

Nozick's truth-tracking account, however, requires that the mechanism track truth in the nearest possible worlds. This is roughly equivalent to tight reasons-responsiveness. Since Fischer only requires loose responsiveness, his account of responsibility is not exactly analogous to Nozick's account of knowledge. He requires just that there is some possible world, which need not be close, in which the agent has reason to do otherwise and acts on that reason.

Fischer (along with Mark Ravizza) criticises Wolf's asymmetry account, arguing that the existence of Frankfurt-style cases disproves her claim that an agent must be capable of doing otherwise in order to be held responsible in cases of blame. Hurley's argument against Wolf is aimed at establishing this kind of point. She argues that we can construct a thought experiment that proves it's possible to generalise Frankfurt's irrelevant alternative intuition, showing that Wolf's requirements for cases of blame *also* regard factors that are irrelevant to whether we should hold agent's responsible.

1.2.3 Susan Hurley's Analysis

Hurley develops her argument by carefully comparing Wolf and Fischer's principles, and separating out the different features of each principle in order to devise

cases in which we can assess independently the contribution being made by each in helping us to establish whether an agent should be held responsible for his act. She hopes to show that Wolf's alternate sequence requirement for cases of blame is irrelevant in much the same way as Frankfurt tries to show that the ability-to-do-otherwise principle is irrelevant in determining whether Jones is responsible.

Hurley notes three major differences between Wolf's account and Fischer's. Firstly, whereas Wolf requires the outright possibility of an alternate sequence for blameworthy acts, Fischer requires just that the actual sequence has certain features. Secondly, Fischer only requires a loose link to reason. If he required a tight link, then those who did not act in accordance with reason would not be responsible. Finally, Wolf requires the ability to respond to objective reasons whereas Fischer's view on this is unclear. This comparison gives rise to three questions:

(1) Does the link to reason needed for responsibility impose alternate sequence demands, or can it be adequately characterised in terms of the dispositions of the mechanisms that operate in the actual sequence?
(2) Is the link to reason required for responsibility tight or loose?
(3) Is it a link to objective reasons, or will subjective reasons do?

Tight reasons-responsiveness, then, is maximal in two dimensions. We can see this when we take both Fischer's distinction and Wolf's distinction into account. It can be weakened either by loosening it (as with cases of weak will) or by subjectifying it.

A deprived upbringing may mean an agent's values are out of line with objective reasons. A person might act on a mechanism that tracks her *subjective* reasons very tightly, but due to her upbringing, these reasons are based on evil values (supposedly, there are no objective evil values). Such a person would be held responsible if we require only responsiveness to subjective reasons, but not if we require responsiveness to objective reasons.

As well as maximal reasons-responsiveness, then, we also have three further categories. We have loose responsiveness to objective reasons, tight responsiveness to subjective reasons, and loose-responsiveness to subjective reasons. These varieties of reasons-responsiveness impose only actual-sequence demands. So long as the agent's actual mechanism is reason-based, it does not matter whether the agent might have acted on some other mechanism instead.

For maximal reasons-responsiveness, however, it's very difficult to distinguish actual-sequence requirements for cases of blame from alternate sequence ones. If virtuous Vivian always acts on a maximally reasons-responsive mechanism, this entails that she will always do the right thing for the right reason. This *a fortiori* implies that she is *able* to do the right thing for the right reason. Likewise, if an agent is not even capable of doing the right thing for the right reason, this implies that she is not acting on a maximally reasons-responsive mechanism.

However, Hurley notes that the implication does not run in the other direction. Salome may be fully capable of acting in accordance with reason, even though the mechanism she actually acts on is not reasons-responsive. There may be the outright possibility of her acting on a different, more reasons-responsive, mechanism instead. But, given the mechanism she does in fact act on, she *would* not do the right thing, even though she *could* do.

In cases where neither Fischer's reasons-responsiveness condition or Wolf's ability condition are met, the agent will not be held responsible. But Hurley wants to work out *which* principle accounts for this. Is it that the agent is incapable of doing the right thing, or that her act results from an insufficiently reasons-responsive mechanism? Hurley suggests we compare cases in which neither condition is met with those in which the ability condition is met, but the responsiveness condition is not. We can then see if the ability condition on its own is enough to make the difference.

In Frankfurt cases it's supposed that the actual-sequence condition is met and the alternate one is not. For these principles, we can't do this. If an agent meets Fischer's actual-sequence requirement (maximal reasons-responsiveness), this implies that they have also met Wolf's alternate sequence requirement (ability to act on reason). Instead, Hurley argues that we can get to the irrelevant alternative intuition by separating the conditions the other way around. We suppose it's the actual-sequence requirement they fail, and the alternate-sequence requirement they meet, and then see how this compares to cases where they fail both requirements. If there seems to be no difference with regards to responsibility, then we will have shown that meeting Wolf's ability condition is irrelevant.

We need to look at examples where, given the mechanism she actually acted on, the agent would not do the right thing even *if* she could have done, and then compare cases where actually she *could* have done with cases where actually she could *not* have done. If the actual causes of her act were such that she would not have done the right thing even if she could have done, is it relevant whether or not it was possible? We can think about this in relation to the three cases in which an agent may fail to do the right thing: Loose reasons-responsiveness (such as weak-will), responsiveness only to subjective reasons (such as evil values), and failure of reasons-responsiveness (as happens with psychiatric disorders). Hurley applies this thought experiment in all three cases.

Wilma is weak-willed – she acts on a mechanism that is only loosely responsive to reason. She would not have done the right thing for the reasons there are, even if she could have (although she might have done if there were slightly stronger reasons). Given she *would* not have done the right thing even if she *could* have, is there a relevant difference between the case in which she could have, and the case in which she could not have? It seems that if this is enough to make her responsible when there is an outright possibility of an alternate sequence, then it's equally sufficient when there isn't.

Ethel is dedicated to evil. If she is not causing any damage, it's because she thinks she can't. If she thought she had the opportunity to do some harm, that's exactly what she would do. She acts on a mechanism that tightly tracks her subjective desire to do evil. Given this, she would not have acted rightly even if she could have, so would it make any difference to how responsible she is if in fact she could have? Again, Hurley argues that the mere ability to do otherwise is irrelevant.

Maude is mentally ill. She acts on a mechanism that is not responsive to reason at all (so of course, she is not responsible for her actions). Because of this disorder, she would not do the right thing even if she could. But if there were the outright possibility of her acting on another mechanism, that *was* reasons-responsive, would this make her any more responsible? It seems once again that if that alternate mechanism played no causal role in her actual actions, we cannot hold her responsible.

The important point in the above cases was that the agent *would* not have done otherwise even if she *could* have done. For Frankfurt's case there is an irrelevant alternative. The reason Jones could not have acted rightly was because Black was ready to manipulate his brain. Hurley argues that the irrelevant alternative intuition is just as valid in relation to determinism – where the agent could not have done otherwise because of the laws of physics and the past. This does *not* mean that she would have done otherwise if she could have done. Even if determinism is true in this world, it might be that, given the mechanism she actually acts on, the agent would have chosen to do exactly the *same* thing had she inhabited a possible world where determinism is false.

Hurley's question of whether an agent would have done otherwise whether or not she could have, relates closely to Frankfurt's question – whether the agent performed the act only *because* she could not have done otherwise. This is the basis of the irrelevant alternative intuition, and Hurley argues that this is equally valid in cases with no counterfactual intervener, where determinism stops the agent from being able to do otherwise.

It is this argument, resting on the cases in which Hurley says we can justify claims of the form "she *would* not have done otherwise even if she *could* have done", when this claim is being made in relation to determinism, which I want to argue rests on a fallacy – the claim that even if we supposed the world to be indeterministic, the actual mechanism on which the agent acts would help us to understand what the agent would be likely to do in this counterfactual indeterministic situation. In the following section, I will try to spell out precisely why I think that this reasoning fails.

1.3 Hurley's Argument & Shope's conditional Fallacy

1.3.1 Hurley's Inference

Hurley's account rests on the claim that in the above cases *given the mechanism the agents actually acted on*, they would not have done otherwise even if they could have done. The idea is that we can infer *from* what the agent *actually* does in an actual situation where the agent cannot do otherwise, what that agent *would* have done had they been in a counterfactual situation where they *were* able to do otherwise. I will argue that this inference is subject to what Shope calls "the conditional fallacy". I will begin by explaining where I think Hurley's principle breaks down, and will then try to formulate precisely why I think it's subject to this fallacy.

She argues that just because in the actual world we could *not* have done otherwise, this does not entail that we *would* have done otherwise had we inhabited a possible world where determinism was false. This seems correct. What is problematic is her claim that it is possible to infer from facts about the actual mechanism, that the opposite *is* true: that the agent *would* have acted on exactly the *same* mechanism had she been able not to. If the first inference is not valid, neither is the second.

In Frankfurt's case there is a clear and simple answer to Hurley's question of whether the agent *would* have done otherwise even if he *could* have. But it's not at all obvious that the intuition can be generalised to the case of determinism, as it's not obvious that such a question can even be answered in the case of determinism.

In Frankfurt's case, it's clear that Black's merely being there on standby makes no difference to what goes on in the example. We can subtract Black from the example without this affecting Jones's actions at all. But it's not at all obvious that we can subtract determinism, and still expect that events would unfold precisely as they would in a deterministic example. Hurley claims that in her examples, the agent would not have done otherwise even if she lived in an indeterministic possible world where she could have done otherwise. It will help to examine exactly what this claim amounts to.

Examining this claim is made tricky by talk of deterministic and indeterministic possible worlds, as determinism and indeterminism are claims precisely about what is or is not possible, and so we end up with modal claims within modal claims. If we are to make sense of this at all, we will have to expand logical space by introducing higher order possible worlds. Purely to simplify terminology, we can call these higher order possible worlds "possible universes". We can suppose there are only two possible universes: one in which determinism is true and one in which it is false.

The universe in which determinism is false will have a course of history that branches off into separate possible worlds at every point at which there is the outright possibility of an alternate sequence of events, keeping all else constant. As a result, the indeterministic universe will have near infinite possible worlds, whereas the deterministic universe will only contain the actual world (assuming determinism is true).

We can also suppose for simplicity that there were only two possible mechanisms the agent might have acted on: The one that leads her to do the wrong thing, and the one that leads her to do the right thing. So, What would it mean for her to have the outright possibility of acting on an alternate mechanism keeping all else

constant? It seems we must say that had she been an inhabitant of the indeterministic universe, it would be equally outright possible that she would come to inhabit either of these possible worlds. There would be nothing stopping her from inhabiting one of these possible worlds over the other, and so it's an open question which mechanism she would have acted on if determinism had turned out to be false. In such a universe, she might act on the mechanism that leads her to do the right thing, but she also might not.

1.3.2 The Conditional Fallacy

It's not at all clear that Hurley's claim – that given the mechanism on which the agent acted, she would not have done otherwise even if she could have done – could have any basis. If the agent's being able to do otherwise entails that both the actual mechanism and the alternate mechanism are *outright possible keeping all else constant*, then it's not clear that there can be any single correct answer to the question of which mechanism she would act on if she inhabited the indeterministic universe.

It may well be that there is no answer to the question of what an agent might have done had she been able to, because if both the alternate sequence, where she acts on the responsive mechanism, and the actual sequence, where she does not, are outright possible, then there simply is no fact of the matter. But even if we do not make as strong a claim as this, it is especially strange that Hurley tries to work out what that fact of the matter might be (if there is one) by looking at the dispositions of the actual mechanism on which the agent acts.

Looking at the dispositional features of the *actual* sequence's mechanism won't help us at all in working out what the agent would have done had it been possible for her to act on a completely *different* mechanism with completely *different* dispositional features. There is no reason why my acting on the mechanism that leads me to do the wrong thing would tell me anything about how likely it is that I would act on an

alternate mechanism if I inhabited a universe where either course of events was possible.

It's true that given the agent acted on an insufficiently reason-responsive mechanism, then *keeping that mechanism constant* he would not have done otherwise even if he could have. But Hurley's examples can only be consistent both with passing Wolf's ability condition, and with failing Fischer's responsiveness condition, if the agent is able to act on a *different* mechanism altogether. This relies precisely on *not* keeping the mechanism itself constant. Wolf's condition will only be passed if it's an open question which of the two mechanisms she actually acts on – if it's an *outright possibility* that she might act on the alternate mechanism.

Hurley's inference from the dispositional features of the actual mechanism to her claim about what the agent would do had it been possible to act on an alternate mechanism cannot be valid. Even if we *could* maintain that there is a fact of the matter about what she would have done had it been possible to act on a different mechanism, whether or not she *would* act on such a mechanism would not be in any way deducible from the dispositional features of the actual mechanism she acted on. So even if there is an answer to Hurley's question, there is no way to infer what it might be.

It might help to consider this in relation to one of her examples. I find it especially surprising that Hurley thinks in Maude's case, we can tell from quite how mad her dispositions are now that she's acting on the basis of a mental illness, anything about what choice she would have made had it been within her power to avoid having a mental illness in the first place, and had she been able to act on a sane mechanism instead. It's hard to imagine how someone could genuinely *be* mentally ill if they were able to *choose* whether or not they felt like acting on a mentally ill mechanism, but we can imagine a scenario in which such a choice is made more plausible.

We can imagine, for instance, that Maude was acting on a reasons-responsive mechanism, until one day she decided to take a tablet that she *knew* would temporarily give her extreme paranoid schizophrenia. She then kills someone, acting on a delusion that was induced by her mental illness. It seems that in this kind of a case, we *do* want to hold her responsible. Perhaps, however, we only hold her responsible because she was acting on a sane mechanism when she took the tablet. But it's hard to imagine any sane person *would* willingly take a tablet that left them with as dreadful a psychiatric disorder as schizophrenia.

So what if we imagine that she did not have any choice in becoming mentally ill in the first place? It seems very strange to suppose that she would have *chosen* to become mentally ill even if she could have avoided it. So in what sense would it be true that she would have done the exact same thing, had she been able to act on a different mechanism altogether? How could we infer from how mad Maude is now that she is mentally ill through no choice of her own, what she would have done had she been entirely able to avoid any mental illness in the first place?

It seems incredible to suppose that the dispositions of her schizophrenic mechanism would tell us anything at all about what she would do if she had had the option of acting on a different mechanism with different (sane) dispositions. In fact, it seems overwhelmingly likely that there are some people who are currently suffering from schizophrenia (and thus acting on an insane mechanism), who would have been willing to take a vaccine to stop them becoming schizophrenic at all had one been offered to them before the onset of their illness. So it's just not true that had they been able to do otherwise, we can infer from the actual mechanism on which they are acting that they would not do otherwise, when doing otherwise would entail acting on some other mechanism.

Firstly, it seems in Maude's case, that it *would* make a difference to how responsible she was if she could have chosen not to act on a responsive mechanism instead. Secondly, it seems, given how unpleasant suffering schizophrenia is, that just

because she actually suffers it, this does not imply that she would have chosen act on the same, schizophrenic, mechanism had she been able to avoid it.

It seems that Hurley's inference commits version 2 of Shope's conditional fallacy. Shope defines this as follows:

> A mistake one makes in analysing or defining a statement *p* by presenting its truth as dependent, in at least some specified situations, upon the truth (falsity) of a subjunctive conditional ∅ of the form: 'If state of affairs *a* were to occur, then state of affairs *b* would occur', when...
>
> ...(*Version 2*) one has overlooked the fact that, in some of the specified situations, statement *p* is actually true, but if *a* were to occur, then it would be at least a partial cause of something that would make *b* fail to occur (make *b* occur).[3]

The following example should help to demonstrate that Hurley's inference is subject to this fallacy:

The government of some country wants to keep its policy of making every individual do ten years of compulsory national service, but it also wants to give the impression that such policies are set in a very democratic way. However, as far as public opinion goes, everyone wants joining the army to be an entirely voluntary matter, and there is widespread outrage about the government's refusal to change their policy in light of the enormous tide of public opposition to it. To get around this problem, the government change the law, and persuade the public that joining the army is now a voluntary matter. They then set up a secret operation whereby every 18 year old who expresses their total unwillingness to ever do anything that might lead them to join the army is spiked with a sophisticated drug that will make them desperately want to spend at least ten years in the army.

---

[3] Shope, Robert K. (1978) "The Conditional Fallacy in Contemporary Philosophy", *The Journal of Philosophy*, vol. LXXV, Number 8, pp. 399-400.

Jim is one of these people who have been spiked with the drug, and the drug makes him act on a mechanism that is not reasons-responsive. It makes him desperately want to join the army despite the fact he has no good reason to. So given the dispositions of the mechanism that he *is* actually acting on, he would not refrain from joining the army even if he could do.

On Hurley's analysis, the statement "given the mechanism on which he is acting, he would not have done otherwise" is analysed as being true on the basis of the conditional: "if he were *able* to act on a sufficiently reasons-responsive mechanism instead, he would not do".

But she has overlooked the fact that if Jim *were* able to act on a mechanism that was responsive to reasons, instead of acting on the basis of this drug, this would cause it to be *false* that he would not do anything other than join the army, since it would cause it to be false that he would act on this mechanism. He only wants to join the army *because* he is acting on this drug. If he had been able to avoid acting on this drug to begin with, he would not have chosen to take the drug or to join the army at all. That is precisely why he had to be spiked with the drug in the first place: because he was unwilling to do *anything* that would ever lead him to join the army. Jim would never have acted on such a mechanism if he had any choice in the matter.

It will help to clarify which elements of Hurley's argument relate to which elements of Shope's characterisation of the fallacy. Shope's statement *p* corresponds to Hurley's claim that because of the actual mechanism on which Jim acts, he would not have done otherwise. The conditional ∅ is the claim that if state of affairs *a* occurs: If he could act on a mechanism whereby he refrained from joining the army, then state of affairs *b* would occur: He would not act on that mechanism and so would not refrain from joining the army.

The fallacy occurs because he would only act on the mechanism that leads him to join the army on the condition that he has no choice in the matter. If he had been able to avoid it, he would have done. So if *a* had been true, *b* would have been false.

Hurley claims that because of the actual mechanism Jim acted on, he would not have done otherwise (statement *p*). The truth of this statement is dependent on the claim (the conditional ∅), that even if it were the case that he *could* act on a mechanism that makes him refrain from joining the army (state of affairs *a*), he *would* not act on such a mechanism, and so would not refrain from joining the army (state of affairs *b*). However, Hurley has overlooked the fact if he could have acted on a different mechanism altogether – one that would not make him join the army (if *a* were to occur), this would make it *false* that he would not refrain from joining the army (would make b fail to occur). *B* would *only* be true if he were acting on the *same* mechanism he actually acts on. But if *a* was true (he *could* avoid acting on the same mechanism) he would *not* act on the same mechanism, and this would cause *b* to be false. (He would not act on that mechanism because he would never *choose* to do anything that would lead him to join the army).

It should be apparent that in cases like Jim's, taking away the alternate possibility does make all the difference. Frankfurt's case is, in important respects, different both to this case, and to the deterministic case, because (unlike the alternate possibilities in these cases) whether we add or subtract Black from the example is not likely to make any difference at all to the way that events unfold, given that it's already specified in the example that Jones's actions are to be kept constant. In Frankfurt's case we alter only the *content* of the alternate scenario, we do not alter how *possible* (or even *probable*) it is that it becomes the actual sequence. This is why Frankfurt's alternative *is* an entirely irrelevant one, but the alternatives that would exist if determinism were false, are not necessarily irrelevant.

Wolf's account concerns precisely the kind of alternatives that are in question when we consider whether the agent's decisions are causally determined. In such cases,

she can hold on to her asymmetry between praise and blame, and she doesn't have to argue that Jones is responsible in Frankfurt's case, because this alternative is irrelevant in a way that indeterminism would not be. So in responding to Hurley's argument, we also see that Fischer and Ravizza's appeal to Frankfurt's case to discredit Wolf's claim that alternate possibilities are required in cases of blame fails too.

1.3.3 Conclusion

It is not obvious that we can generalise the irrelevant alternative intuition to help us with the case of determinism, as it poses distinct problems to those accounted for in Frankfurt's case. Hurley's analysis brings these problems out more clearly. It does not help us to resolve them. For this reason, Wolf's account cannot be so easily discredited, and I think that her suggestion that responsibility requires the ability to be determined by objective reasons should be considered in more detail. In considering the other features of her account more carefully, we might become better equipped to answer the question of whether responsibility requires actual or alternate sequence principles.

We cannot answer this question with reference to Frankfurt examples alone. I want to suggest, instead, that we should look carefully at what it is we take to be of moral significance when we make attributions of praise and blame, and why we consider it important for agents to respond to their moral obligations.

If we want to assess whether Wolf's account works, rather than considering its relation to Frankfurt's case, we might want to look at the kind of metaethical claims that Wolf's account commits us to. It seems reasonable to suppose that a plausible account of responsibility should relate in some intelligible way to other fundamental issues concerning the rightness or wrongness of acts.

The relation between responsibility, reasons, and objectivity plays a large role in Wolf's account. If we want to consider the plausibility of Wolf's conditions, we should at least be able to say something about how we take these issues to be important.

Over the next three chapters, I want to look at the role of reasons and objectivity in supporting Wolf's account of responsibility, and at the kind of problems we encounter in trying to develop a plausible understanding these features. Having tried to formulate a good understanding of these issues and the distinct problems posed by them, I will then return to the specific issue of moral responsibility, and will consider what kind of an account of the conditions of responsibility we might end up with, taking these further conclusions into account.

In the following chapter, I will look specifically at the problem of how we could have objective reasons of the kind Wolf's account needs – reasons that could apply to us regardless of our subjective desires. This claim is not an uncontroversial one, and there are various grounds on which other theorists have sought to reject it. I will try in the following section to address the kind of argument that is often given for this sort of conclusion, and to argue that there could exist reasons of the kind that Wolf's account requires, so long as we adopt the correct analysis of certain types of reasons claim.

2 Rationality & Responsiveness to Reason

The previous chapter was aimed at defending Wolf's Reason View against the kind of argument (put forward by Fischer and Ravizza, and more carefully by Hurley) that tries to discredit the account by drawing on Frankfurt cases and the "irrelevant alternative intuition". I tried to establish that such considerations in themselves are not enough to discredit her account.

Part of my reason for defending this particular account (or at least certain aspects of it) is that many of Wolf's insights into the issues surrounding attributions of responsibility seem to me to be especially valuable, and seem to take us a lot closer to the relevant considerations when it comes to accounting for our intuitions about responsibility. It allows us to distinguish quite accurately between the kind of cases where we think such attributions are justified, and the kind where we do not.

There seem to be two fundamental insights driving Wolf's account. Firstly, she does not think that freedom in itself is what we are concerned about in terms of accounting for responsibility, but rather that the agent needs to have the right kind of capacity for doing what's right, hence the fact that we end up with an asymmetry between cases of blame and praise. If the agent manages praiseworthy action in the first place, then we already know that agent has the capacities necessary for it. Secondly, Wolf is driven by the idea that we need some standard of judging what is objectively valuable in order to hold agent's responsible for their actions.

In relation to this last point, Wolf often talks about "the true and the good". The idea that some act can independently have or lack value is considered a precondition for having any genuine basis on which to consider actions to be praiseworthy or blameworthy. It's often acknowledged in discussions of freedom and responsibility, that if we lacked responsibility this might threaten our concept of moral

values. But the fact that a lack of any independent basis for moral values might threaten the concept of responsibility is rarely discussed. On Wolf's account, the reasons we have for being moral in the first place are closely linked with the reasons we have for considering an agent to be blameworthy or praiseworthy for performing an act.

The importance we place on the agent's reasons for doing the right thing in the first place seems to be a large factor in our holding that agent responsible for his act. The idea that responsibility naturally requires that there is a true and a good that agents are expected to respond to, seems to at least be a worthy suggestion, but also a controversial one. For both of these reasons, I think this aspect of Wolf's account needs to be considered in more detail. Although the idea has rarely been put forward, it seems a plausible supposition that our ability to account for whatever it is that grounds the goodness of an act should be important in terms of establishing why that act might be considered praiseworthy.

Furthermore, if we can develop principles of responsibility that reflect a plausible understanding of why certain actions might have the value we attribute to them in the first place, we will have an account that actually relates to the kind of metaethical considerations that explain our reasons for doing the right thing to begin with. In this respect, our account of responsibility will also benefit from being more *to the point* when it comes to considering why an agent's decision matters to us in a way that would lead us to make the kind of judgements associated with praise and blame.

In this respect, Wolf's account's dependence on there being objective reasons for action seems to be a good thing. But such a claim is also controversial. It has often been argued that reasons simply are not the kind of thing that can *be* objective in the way that Wolf's view requires them to be. If the existence of objective reasons is going to be a fundamental part of the way we account for responsibility, then we should at least hope to show that reasons of this sort are able to exist at all.

I will try to defend Wolf's account against the kind of argument that's sometimes put forward in favour of the conclusion that there could be no reasons that are "objective" in the sense that Wolf's account requires. I will focus on Bernard Williams's famous discussion of internal and external reasons. I aim to defend what I take to be the most significant aspects of Wolf's account in light of this kind of argument.

I will begin by looking at exactly what the claim that we must be able to respond to objective reasons amounts to. I will then look at what grounds we might have for rejecting a theory that requires reasons of this sort.

2.1 Objectivity, Rationality & Responsibility

2.1.1 Objective Reasons & Rationality

One immediate problem we encounter in trying to clarify exactly what it is to respond to objective reasons is that there is very little agreement among philosophers of precisely what constitutes a reason. The most basic feature we might expect a reasons claim to have, is well summed up by Thomas Scanlon, when he says that a reason should "count in favour" of something[4]. Even this basic condition is not accepted on all philosophical accounts, but it should be uncontroversial enough that we can take it for granted for the purposes of this discussion. As soon as we get beyond this very basic feature, however, what we mean is likely to be significantly more debatable.

On some models, a reason is taken to be the kind of thing that we can cite as the cause of an agent's actions (at least in conjunction with certain facts about the way humans operate). It's sometimes taken to involve means-end reasoning on the basis of

---

[4] Scanlon, Thomas. (1998), "Reasons" in *What we Owe to Each Other* (Harvard University Press) p. 17.

beliefs and desires. But we also think of reasons as the kind of thing that we could offer someone as advice, assuming that so far they are playing no causal role in determining their actions, and might never do so, in which case the force of such reasons is taken to lie just in the fact that we think they *should* be.

So reasons not only tend to have an explanatory role in making sense of a person's actions from an outside perspective, but for the agent to whom they apply, we expect that reasons will have a certain kind of prescriptive or normative force. Acknowledging the existence of a reason is supposed to affect the agent's motivations.

This is all consistent with Scanlon's basic definition, but we might be tempted to think we can say more. It seems like what we have reason to do is explained by what course of action is *rational* for us. If we want to understand the significance of the claim that an agent must be capable of responding to objective reasons, we must, at the very least, look more carefully at the relation between reasons and rationality.

It seems like an obvious supposition, in light of the above considerations, that reasons are explained by rationality. We might suppose that if some consideration counts as a reason for an act, this will be *because* of the fact that it is rational for us to perform that act. This seems to explain the kind of normative pressure we feel to shape our intentions in accordance with the reasons that exist.

This link between responding to reasons and being rational certainly seems to form a large part of the way that Wolf understands the force we take reasons to have. In arguing against the view that we should require alternate possibilities as a condition for responsibility in cases of moral praise, she often asks about what it would mean to be "able to *not* do the right thing for the right reason", and here she uses this phrase interchangeably with "able to act irrationally". Likewise, she often uses the term "able to do the right thing for the right reason" interchangeably with "able to act rationally". This indicates that she understands what it is to have a reason for some act in terms of what it means for a certain course of action to be rational for an agent.

It seems reasonable to suppose then, that Wolf sees the force of reasons as being explained by what is rational. This is one claim we might want to consider more carefully in evaluating Wolf's account.

2.1.2 Reasons & Objectivity

So far, I have been talking about what it is to have a reason generally, and about how Wolf understands this. But it might seem like a more controversial or objectionable feature of Wolf's account is related, specifically, to the notion of having an *objective* reason.

Wolf defines "objective reasons" as implying "...the existence of non-arbitrary standards of correctness, standards which are independent of an individual's will and even of an individual's psychology as a whole, by which one can judge some actions, choices, ways of life, or systems of value to be better than others"[5]. This *independence* of the reason from the individual agent is the basis on which such reasons are considered to be objective.

As noted in the previous chapter, Wolf draws a sharp contrast between the Reason View and the Real Self View. She objects to the Real Self View on the grounds that merely being able to respond to your own subjective system of values is not enough to ground responsibility. In order for an agent to be responsible, her subjective values must also be determined by the true and the good.

In cases where an agent has a very bad upbringing, his subjective values and motivations might differ drastically from the true and the good. This might lead the agent to have subjective reasons that differ wildly from the *actual* reasons he is

---

[5] Wolf, Susan. (1990) "The True and the Good" in *Freedom Within Reason* (Oxford University Press) p.124.

objectively subject to. E.g. given that an agent has evil values, this will give him subjective reason to perform actions that will cause harm, as this will best further those values that he actually accepts. Objectively, however, he will have no such reasons because this system of values itself is actually worthless. It does not coincide with the reasons there really *are* independently of any subjective aims or motives the agent might have.

2.1.3 The Cause for Objection to Wolf's Account

In order to see why Wolf's claims might be considered controversial, it will help to get clear about what she is actually committed to. Her account seems to have two main threads to it. On the one hand, the force of reasons is explained in terms of what is rational for us, and on the other hand, reasons are considered to be "objective" in that they do not depend on the agent's subjective aims or desires. So we have the following two claims:

(1) If an agent has a reason to act, this will be explained by the fact that such a course of action is rational for the agent.

(2) There are reasons for action that are objective in that they are independent of the agent's subjective values.

On the face of it, it does not seem like there is any tension between these two claims, but this will depend on how we understand the term "rational" as applied to action. There is a plausible account of what it is for a course of action to be rational for an agent, spanning from Hume and defended by Williams, whereby a course of action will only be rational for an agent on the basis of certain subjective aims or values that agent already has.

Specifically, Williams argues that an action is only rational on the condition that we have some subjective motive that would be furthered by it (in Williams's terminology, if there is a "sound deliberative route" leading from some aim or motive that the agent cares about to the conclusion that there is something to be said for performing the action).

If an action only ever counts as rational because of it's relation to the agent's values, then clearly reasons cannot be objective (in the sense of being independent of the agent's values) *as well as* being explained by what is rational for the agent. Yet Wolf's account seems to require both. Williams's account requires that we give up claim (2). I want to suggest we give up claim (1) instead.

Ultimately then, I do not want to defend every aspect of Wolf's account. But I do want to defend a significant portion of it. In particular, I do think that responsibility requires objective reasons, and that there can *be* objective reasons in the sense that Wolf's account requires, i.e. an independent standard by which an agent's subjective values can be assessed as good or bad. This objectivity, however, commits Wolf to reasons that are also "external" as defined by Williams. They do not depend on any subjective aims or motives the agent might have.

Williams argues that no reasons of this kind could possibly exist because the existence of such reasons would not imply that such courses of action were rational. If however, we do not think reasons are *explained by* rationality, this will not be a problem.

In the following section, I will look more carefully at Williams's account of rationality in action, and at how he aims to establish that we could not have external reasons.

2.2 Internal & External Reasons

2.2.1 Williams's "Sub-Humean" Model

Williams's view of practical rationality is much broader than the Humean model that it's taken to be an adaptation of. Hume only allowed practical rationality to cover instrumental means-end reasoning. Desires are not beliefs about matters of fact, and so cannot be judged in terms of how rational they are. Courses of action can only be considered practically rational insofar as they constitute a means to satisfying some desire the agent already has.

Williams's model is based on this Humean model, although it's considerably less narrow with regard to what elements count as fulfilling the function of a desire, and with regard to what counts as reasoning about how to satisfy these elements.

Whereas Hume included only desires in the classic sense among those elements that could give rise to a motivation, Williams says that any elements in an agent's "Subjective Motivational Set" (an agent's "S") can be the basis for a reason, and this can include "such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent". He also adds, "Above all, there is of course no supposition that the desires or projects an agent have be egoistic; he will, one hopes, have non-egoistic projects of various kinds, and these can equally provide internal reasons for action".[6]

Likewise, whereas Hume included only reasoning about the means to a given end that the agent desires, in terms of what could make it rational for an agent to perform some action, Williams argues that all we need is a "sound deliberative route"

---

[6] Williams, Bernard. (1981) "Internal and External reasons" in *Moral Luck,* (Cambridge University Press) p. 105.

leading from some element in the agent's S to the conclusion that there is something to be said for performing the action, and he defines having a sound deliberative route very broadly.

He argues that "a clear example of practical reasoning is that leading to the conclusion that one has reason to ϕ because ϕ-ing would be the most convenient, economical, pleasant, and so forth, way of satisfying some element in S, and this of course is controlled by other elements in S, if not necessarily in a very clear or determinate way. But there are much wider possibilities for deliberation, such as: thinking how the satisfaction of elements in S can be combined, for instance, by time-ordering; where there is some irresoluble conflict among the elements of S, considering which one attaches most weight to...or, again, finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment"[7].

Williams is willing to allow for the fact that an agent may be *unaware* of the fact that some course of action will satisfy a desire, or the agent may only have certain elements in her S on the basis of false belief, and this may hinder her actual ability to deliberate to the proper conclusion about which acts there is something to be said for performing. Ignorance of these facts, or even ignorance about elements within the agent's S (perhaps because some elements are in the unconscious), may mean the agent is *unaware* of the fact that a sound deliberative route exists. But if finding out certain truths would make the agent see that there is such a reason (so long as she deliberates rationally), we can say that there *is* a sound deliberative route, and it's just that she doesn't know about it.

The main thread linking Williams's view to Hume's is his requirement that all genuine reasons claims will have subjective conditions – and therefore that they will be *internal* reasons claims. An agent will only have a genuine reason to perform some act

[7] Williams, Bernard. (1981) "Internal and External reasons" in *Moral Luck,* (Cambridge University Press) p. 104

if there is something that matters to him relative to which such an act is rational. So the agent's having some element in his S; some desire, aim, motive, or value from which there is a sound deliberative route to the conclusion that there is something to be said for performing an act, will be a *necessary condition* of that agent having a reason to perform the act. So, any internal reason statement will be falsified by the absence of appropriate elements in the agent's S.

2.2.2 The Sub-Humean Model & the Falsity of External Reasons Claims

Williams argues that it's unintelligible to suppose that an agent has a reason to perform an act unless there is a sound deliberative route leading from some element in the agent's S to the conclusion that there's something to be said for performing that act.

He offers two considerations in support of this point. Firstly, he argues that we cannot explain the way reasons motivate us without supposing this kind of a rational relation between such reasons and elements in the agent's S, and secondly, he argues that when this kind of rational link is lacking, to tell an agent that they have a reason to perform such an act amounts to mere browbeating.

Williams argues that external reasons theorists will have serious theoretical difficulties when it comes to explaining how it is that an agent comes to form new motivations in response to acknowledging a reason. If there is no appropriate element in the agent's S that could rationally support such a motivation, it seems impossible to explain how that agent ever reaches the conclusion that there is something to be said for performing the act.

Williams expects that we might be tempted to think that coming to believe that an external reason statement is true will help explain an agent's acquiring the new motivation, but he argues that this *in itself* will not explain anything.

Williams asks "*What* is it that one comes to believe when he comes to believe that there is reason for him to ϕ, if it not the proposition, or something that entails the proposition that if he deliberated rationally, he would be motivated to act appropriately?".[8] If this is what such an agent comes to believe, then on Williams's view, this entails believing that there is some appropriate element in his S from which there is a sound deliberative route to the conclusion that there is something to be said for ϕ-ing.

So in order for an external reason statement to explain how the agent forms a motivation, it must already be true that the agent has some appropriate element in his S of the kind that would justify an *internal* reason statement, and it would be this internal reason statement that was doing all the work in explaining why the agent came to form the motivation.

This point about motivation is one of the ways that Williams supports his position. The other is to argue that telling an agent that he has a reason to perform an act when he has no appropriate element in his S from which to deliberate to such a conclusion amounts to mere browbeating. To demonstrate this, Williams uses the example of Owen Wingrave.

Owen Wingrave's father wants Owen to join the army, and insists that their long-standing family tradition of military honour is a reason for him to do this. Despite knowing there is absolutely nothing in Owen's S that would ever lead him through any amount of deliberation to join the army, his father insists on saying that this is what Owen should do.

---

[8] Williams, Bernard. (1981) "Internal and External reasons" in *Moral Luck,* (Cambridge University Press) p. 109

If it's true, as Williams argues, that we will not have any reason to perform an act, unless we have some basis in our S from which we can deliberate to the conclusion that there's something to be said for it, then we would expect that telling someone to pursue a course of action for which there is no basis in their S would amount to mere browbeating. The case of Owen Wingrave serves to confirm that this is precisely how such a situation strikes us. Owen's father isn't appealing to anything that Owen can be expected to respond to in virtue of being rational and understanding the relevant considerations. Given this, it just seems plain false to suppose that Owen should join the army.

This example has some obvious flaws. It involves an entirely implausible reason, and so we might think it's the fact it's implausible, rather than the fact that it's an external reason, that accounts for its falsity. Supposedly this will not be the kind of thing that external reasons theorists have in mind. I will later discuss whether using a more plausible reason would make the browbeating claim lose its force.

2.3 External Reasons, Values & Moral Obligations

2.3.1 Supporting External Reasons in Light of Williams's Argument

It seems that if we are going to accept these arguments, and accept Williams's claim that such considerations do support his main point (that all genuine reasons have subjective conditions) we will also need to accept some background assumptions.

Firstly, we have to suppose that it's not possible for an agent to form new elements within the S on the basis of rational deliberation. There is an assumption (spanning from Hume) in Williams's argument that desires are the kind of thing we are merely subject to, and cannot be assessed in terms of how rational they are.

Secondly, we must assume that we will not have a reason to perform any act, unless that reason is explained by the fact that it is rational to perform that act. We could not have a reason to perform an act on any basis other than the fact that such a course of action is rational.

We might, however, have grounds for rejecting both of these assumptions. To see why this is so, it will help to consider some of the points raised by Scanlon in his discussion of Williams's argument.

Scanlon points out that Williams is willing to concede that an agent whose S fails to support certain reasons may be justifiably subject to a range of criticism, so long as we do not accuse such a person of irrationality.

Williams says: "There are of course many things that a speaker may say to one to is not disposed to ϕ when the speaker thinks that he should be, as that he is inconsiderate, or cruel, or selfish, or imprudent; or that things, and he would be a lot nicer if he were so motivated. Any of these things can be sensible things to say. But one who makes a great deal out of putting the criticism in the form of an external reason statement seems concerned to say what is particularly wrong with the agent is that he is *irrational.* It is this theorist who needs to make this charge precise: in particular because he wants any rational agent, as such, to acknowledge the requirement to do the thing in question".[9]

Scanlon points out that if we can justifiably aim any criticism at the agent at all, we will be implying that their failure to respond to a particular reason is a *deficiency* on the part of that agent (albeit a non-rational deficiency). If failing to see something as a reason is a deficiency, this implies that it *is* genuinely a reason. If no reason actually existed, we would have no justification for considering a person to be guilty of *any* kind of a deficiency when they fail to respond to it.

---

[9] Williams, Bernard. (1981) "Internal and External reasons" in *Moral Luck,* (Cambridge University Press) pp. 110.

Having established this, Scanlon infers that even Williams himself must be committed to the claim that any agent who does not see certain considerations as a reason is deliberating in a *faulty* manner. Just taking into account the fact that we can criticise an agent for failing to see something as a reason, we can establish that it is a reason, and from this we can conclude that if an agent fails to see it as a reason, they cannot be thinking about the matter correctly.

But even given that an agent who does not see some consideration as a reason must be subject to faulty deliberation, we still have a problem when it comes to working out what it is that the agent is expected to deliberate from in order to reach this conclusion. Scanlon points out that Williams allows that the process of deliberation may affect the contents of an agent's S in various ways. We can imagine that a man who is bored on his weekends has a reason to take up some kind of an interest, and that this would then give rise to various other aims and desires that he lacked before. To some degree then, even on Williams's view there is a commitment to the claim that the elements contained in the S can be subject to rational deliberation.

Given this, Scanlon points out that if Williams is willing to say that a person is deliberating wrongly who fails to reason from some element in the S to the conclusion that there is something to be said for performing some act, it seems odd that he is not willing to accuse an agent of defective deliberation if the reason they do not pursue some course of action is because they fail to learn to value of it the first place, given that they do have an independent reason to value it (which they must do if *any* criticism against an agent who fails to see the value in such things is justified).

2.3.2 Reasons & Value

It seems true that if we can criticise an agent on the grounds that his S does not support some reason, and if such criticism is going to be justified, we might think

this implies that there is genuinely a reason to perform the act. However, the fact we are justified in calling someone "unpleasant" or "cruel" or "inconsiderate" does not necessarily imply that there is any genuine reason for that agent *not* to be any of these things. It's merely the fact that such terms are considered *critical* of the agent that implies that they are genuine flaws, rather than merely features of that agent we personally dislike (in the same way we might dislike their height). Williams might not accept this, but I take it that most of us will find it harder to deny that such features are genuine *flaws* in a person's character.

Scanlon's argument begins just with the assumption that there is something that's actually *valuable* about not being inconsiderate, cruel or selfish. There is at most a very weak supposition that this is *because* it's irrational to be cruel, inconsiderate or selfish.

In the following section, I will argue that the attempt to ground these kinds of values in some concept of rationality is a step that we do not need to take in order to get around Williams's argument.

I will try to offer a way of supporting the claim that there can be objective (and thus external) reasons. But I will do this only at the expense of one of the claims that seemed to be playing a large role in Wolf's account – the claim that the force of such reasons can be directly explained by which courses of action are rational for an agent. I want to argue that the relation between rationality and certain reasons claims will have to be construed as more complex than this.

This particular construal of the relation between rationality and reasons that we find in Wolf does not seem to play any important role in supporting her account of responsibility. That depends just on there being *objective* reasons, and we can accept this claim without having to buy into the claim that reasons are always explained by what is rational.

We could get around Williams's argument, even if we *accepted* that agents only have reasons based on their ability to rationally deliberate to some conclusion in favour of performing an act. We could do this by arguing that we could rationally deliberate to conclusions about what it is we should *value* to begin with.

This is a possible approach, but it's not one I want to attempt. I think it's at best very difficult to sustain the claim that we have rational deliberative control over all of our motives, aims and desires. I don't intend to say anything that would rule this claim out, but I don't need to support it either. For the purposes of this discussion, then, I will simply take for granted Williams's view that we cannot expect all genuinely rational agents to have the correct motives to begin with.

Williams's argument also rests on our assuming reasons are always explained by what is rational for an agent. This would entail that if there were objective reasons, all rational agents would be motivated to respond to them. However, I want to reject the claim that reasons are always explained by what is rational. Without this claim, Williams's argument breaks down. My motive for rejecting this claim is that I think there are certain reasons for which such an explanation is not plausible.

2.3.3 Reasons, Rationality & Moral Duties

I want to argue that whilst Williams's view might be plausible in accounting for some reasons, there are others that it does not seem to account for convincingly. *Some* reasons do seem to depend on the agent having particular motives that bear a certain rational relation to some end. But it also seems there are nonetheless cases where without any such motives we *do* still think that the agent has a reason.

It will help to look at the different grounds on which we might find Williams's account implausible. In doing this, we might gauge a better idea of where the account derives its plausibility and what the limits of this are.

On the one hand, we might think that the whole notion of rationally deliberating on the basis of the contents of the agent's S seems to involve a lot more cognising about means and ends then we ever generally do when we are thinking about what reasons we have. An agent might see the fact that his library card has expired as a reason to renew it, and he might consider it a reason without ever introspecting about the contents of his S in order to determine whether he genuinely has any motives in light of which such an action would be rational.

The problem is, whilst this may show that there is some problem with the requirement that we are *aware* of a sound deliberative route from some element in the S, it does not seem to undermine the claim that (whether we think about it or are aware of it or not) it will only be *true* that we have a reason to renew the library card on the condition that we have some element in the S that would be furthered by this.

To say that an agent will only have a reason on the condition that he has some appropriate element in his S is not to make a very bold claim – it does not say anything at all about what that agent will actually consider whilst he decides what he has reason to do. It just means that if there really is *no* such element, then the reasons claim will be false. E.g. if he does not wish to return to the library, has no commitments to any other institution he cares about that require him to have an up-to-date library card, does not value the abstract possibility of library access in any way, and/or has two hours to live and more important things to do than read, etc., this *does* plausibly falsify any claim about him having a reason to renew his library card, and this seems to show that such conditions really are necessary for the existence of this kind of reason.

If every basis we could possibly imagine within the agent's S that would make such a course of action rational was lacking, it seems that it would be true that the agent has no reason to renew his library card. None of this involves making any claims at all about what the agent is likely to actually think about when he is deciding what he has reason to do.

The claim that having some appropriate element in the S is a *necessary condition* of having a reason for a given course of action, and therefore that the absence of any such element will falsify a reasons statement, seems to me to be plausible and weak enough that it's difficult to deny it for many cases. I cannot help but find this very plausible for the above case, although I am sure there are theorists who don't even find it plausible for this example. I personally find that surprising, but at any rate, I do not intend to discuss the issue of why even in this case some people might not be convinced. This is because I am not worried about what plausibility it may or may not have for *this* kind of case, because I don't think that it maintains its plausibility for *every* case, even when understood weakly as just a claim about the necessary conditions for having a reason.

The cases in which I think the explanation breaks down, in particular, are cases involving moral reasons.

2.3.4 Moral Reasons

Scanlon offers the example of an agent who is cruel to his wife, and is surprised that Williams holds a view that involves claiming that such an agent will only have reason to treat his wife better if there is something he values in light of which it would be rational to do so. He argues that this would force us to the conclusion that our own reasons for refraining from being cruel to our spouses are also based on subjective conditions, and that this rings false.

The important point in this case, that we might think was lacking in the library card case, is that it seems to ring false that such an agent has no reason to treat his spouse better *regardless* of what elements we suppose his S to contain.

Whereas in the library card case, when we imagine that all subjective conditions are lacking, it seems impossible to deny that this falsifies the reasons claim, this doesn't seem to be true in the current case. We can imagine that he does not like his wife very much, that he does not value being kind to people, that he enjoys treating her with his current level of cruelty, and that he is pretty certain that she's now too emotionally weak to ever phone the police and file a complaint or anything like that, etc. Whilst this might make us predict that it's unlikely he *will* start to treat his wife better, none of this seems to make us conclude that there is really no *reason* for him to treat his wife any better.

In this kind of case, it's Wolf's claims that seem to accord especially well with our intuitions. Our inclination is precisely to think that even if he has subjectively based reasons to be cruel to his wife, this would not show that he has no reason to treat her better; it would simply show that he has reason to reject his subjective system of values. Any equivalent claim we might be tempted to make for the library card case would be completely implausible. Unlike with the case of not being cruel to our spouses, we think that the reasons we have for renewing our library cards *are* completely exhausted by considerations that depend on our subjective aims and motives.

It seems that the main difference between these cases is that one of them involves a *moral duty*. We do not tend to think that what we have a moral obligation to do is determined by our subjective motivations in the same way as other reasons are. Whereas we tend to think that a claim about our having reason to buying a new television would be falsified by a total lack of subjective conditions, we do not think that our reasons for not mugging old ladies would be. If we do not care about the welfare of other people, this will not indicate that we don't *have* a reason; it will indicate that we are not *acknowledging* the reasons we have.

Furthermore, whereas telling a person who has no basis in his S that he should renew his library card or join the army might seem like mere browbeating, we do not

think that telling an agent she should stop abducting people at gunpoint and keeping them as prisoners in the attic counts as browbeating in the same sense. If an agent is in the habit of holding innocent victims hostage in the attic for her own amusement, we would not ask her to rethink this policy *because* we think such acts might not be rationally supported by her subjective values. We just think it's wrong full stop, whether or not she values anything that would lead her, in virtue of her rationality (or anything else for that matter), to think it's wrong.

I would also like to argue that we do not think such actions are wrong *because* we think they are irrational. We need not claim that every rational agent will respond to such considerations purely in virtue of being rational. We might just think that the very fact it's morally wrong (irrespective of whether it's rational) is a reason – and one that's there objectively – whether or not anyone cares about this fact. This indicates that we think there is something more that accounts for the force of moral obligations.

If we think that an internalist perspective on reasons works for most reasons claims, but breaks down when we think about moral reasons, we will nonetheless have established that there are objective reasons of the kind Wolf's account requires for responsibility, for the cases that are arguably most interesting and important. We will be able to formulate a good account of *moral* responsibility.

2.3.5 Moral Obligations & Rationality

Ultimately, I will try to argue that the difference between the case of moral reasons and the case of non-moral reasons is to do with the fact that the normative force of reasons claims is best explained differently in the moral case to the way it is best explained in most[10] non-moral cases. To the extent that we might think it relates

[10] I say "most" non-moral cases because we might think that we have obligations that do not rest on moral considerations. E.g. we might think we have reason not to destroy great pieces of art, even if no one would know about it or be hurt or upset because of it. This equally will rest

to rationality, it certainly does not relate in the same way we might take other reasons to.

Most of the things we consider ourselves to have reason to do are not driven by moral considerations, and so we tend to think that such reasons are based on what is rational alone. Furthermore, it seems likely that however we understand the notion of rationality, we consider it to be what is rational that determines most of the reasons we have, rather than vice-versa. But the case of moral reasons is different. We think that the moral wrongness of an act is a consideration that imposes normative pressure on us in a way that's largely unrelated to the kind we associate with thinking about whether a certain course of action is rational given various aims or desires we have.

Part of the problem with Williams's argument is to do with his assumption that we will not have a reason for action unless there is this particular *kind* of rational pressure to comply. This relates to his supposition that there is no sense in which we might say that someone has a reason, and not mean by this that it might play some direct role in explaining that person's motivations. This leads him to assume that anyone who wants to defend external reasons will want all rational agents to respond to the requirement in question purely in virtue of being rational.

If this is what we want to claim, then we probably *do* have a problem when it comes to explaining how an external reason statement all on its own is going to give rise to the right motivation, since such a claim implies that acknowledging the reason and thinking about it rationally will automatically provide the motivation we are looking for. But if we do not want to make this kind of claim, we do not run into this problem. We do not have to suppose that the mere existence of a reason implies that

---

on thinking that art can be *objectively* valuable in some respect. Whether or not this kind of case is plausible is largely outside of the scope of what I want to discuss, but it's worth noting that nothing I have said should rule out the possibility of objective aesthetic values of this kind counting in a similar way to moral values.

any rational agent will *necessarily* be motivated to comply with it, and we can still maintain that such a reason exists.

This is a very important point when it comes to considering exactly how we might think the notion of rationality actually relates to moral normativity. There are various ways we might define the concept of "rational", but if we think that some consideration can motivate all rational agents in virtue of their rationality alone, and we think that it's in *this* sense moral reasons depend on rationality, we will have to do a lot of work in order to avoid falling prey to Williams's criticism.

Once we have moved from the problem of reasons claims more generally to the specific issue of moral reasons, this problem of how an external reason can give rise to a new motivation for any agent purely in virtue of their rationality is often conceptualised as the problem of how we can explain the force of moral reasons in such a way as to convince even an amoralist who does not care about morality, just in virtue of their rationality alone. If, however, we do not think that the normative force of rationality directly explains the normative force of moral reasons, we have no reason to require that anything like this is possible. We can say that any *moral* person would respond to a consideration in virtue of being moral, without saying that every *rational* person would respond to it.

At this point, it might be useful to mention some considerations on this subject raised by Joseph Raz, in his discussion of Frankfurt's writing on respect. Frankfurt argues that we should treat people with respect, rather than treating them equally, in the sense that we should only discriminate between them on grounds that are relevant. He notes that this seems to be a rational requirement, but that it can't be important just for that reason, since acting irrationally, does not imply acting immorally.

Raz argues that we might consider it irrational to fail to treat someone with respect purely on the basis that we consider it irrational to fail to act on an undefeated reason, and we have an independent moral reason to treat people with respect. In this

case, he points out that it will be true that we *could* accuse someone of irrationality when they fail to act on such a reason, but to do so would, if nothing else, completely miss the point. The basis we have for criticising someone is independent of any irrationality they might also be guilty of as a result, since the force of their reasons for treating others respectfully is not itself *explained by* the fact it would be rational. Instead, the fact they would be acting irrationally by failing to do this would be explained by the fact that it's wrong, not vice-versa.

Clearly, moral reasons are associated with a normative pressure to respond, and so are reasons that arise purely out of rational concerns (such as means-end reasoning or good belief-forming practice). But the demands of morality do not seem to get their force in virtue of the demands of rationality. Clearly there is normative pressure to comply with our moral obligations, but we will have to explain this pressure in some other way. We cannot explain why it's important not to be cruel to our spouses using the same considerations that would explain why we have a reason to renew a library card.

In the following chapter, I will attempt to make clear precisely where I think the force of moral reasons claims lies. In doing so, I will try to work out exactly where the boundaries are between performing an action because it is rational to do so, and performing an action because of a moral duty to do so.

The account I wish to defend is broadly Millian, in that it involves drawing a link between the ways in which people would be justified in responding to an act and the reasons we have for performing or that act. This account does not entirely divorce the notion of moral reasons from that of rationality, but it is nonetheless consistent with accepting some of the persuasive anti-rationalist points that span from Humean theory, which have influenced some of the considerations that have been the focus of this chapter.

Taking into account the conclusions of this chapter, and applying this to Wolf's account, we end up with a model where responsibility rests on ability to respond to objective reasons, but unlike on Wolf's account, we do not take those reasons to be directly explained by rationality. In the following chapter, I hope to support an account of the way we should explain the force of moral reasons that relates in a more direct way to the justifications we might have for the kind of attitudes we associate with blame and praise. As mentioned earlier, if we can do this, the idea that responsibility relates to reasons will seem significantly more *to the point* when it comes to accounting for moral responsibility and it's significance.

3 Moral Obligations, Reason & Normativity

I will now try to clarify how the current points fit in with the overall aims of this discussion. I began by trying to support Wolf's Reason View of responsibility, and by noting that it cannot be discredited as easily as some theorists have thought. But this view does rest on certain claims that might be considered objectionable on other grounds. It's committed to the existence of objective reasons.

For reasons to be objective, they must be independent of an agent's subjective values and desires. Williams argues that reasons of this kind cannot exist. It seems, however, that Williams's argument only works on the assumption that all reasons are directly explained by rationality. He supports this assumption by noting that if someone lacks the right kind of a rational basis for a reason, we cannot explain how he comes to be motivated by it, and telling him to act on it amounts to browbeating.

For many types of reasons claim this might seem plausible, but for moral reasons this explanation breaks down. Unlike with other kinds of reason, if we think someone has a moral duty to do something, we need not suppose they will be motivated by their mere acceptance of this fact, and it does not seem to be browbeating in any relevant sense when we tell someone to acknowledge a moral obligation.

But this means we drop the assumption that reasons have their normative significance in virtue of considerations about what is rational for an agent to do. We need to suppose that the force of such reasons is explained by something else.

In this chapter, I will argue that we can account for the force of moral reasons in terms of considerations about the kind of attitudes that seem appropriate in response to certain acts. If we explain moral reasons in this way, we will have an

explanation that's much more to the point when it comes to understanding why we might suppose there is the kind of close link between reasons and responsibility that we find in Wolf's account.

The considerations in this chapter, should relate the way we explain moral reasons claims directly to considerations that are relevant to the issue of responsibility. This would not be possible if we wanted to explain reasons in terms of which courses of action are rational for an agent. This is because the kind of normative significance we associate with moral duties seems to support certain attitudes in response to certain acts, and this fact cannot be accounted for with reference to what actions are rational from the agent's perspective alone.

I hope to show that once we start to think of moral reasons in this way, we will be better equipped to return to the initial problem of freedom and responsibility, and to evaluate precisely what it is that accounts for the important role reasons seem to have in supporting an account of responsibility. We will then be able to reconsider the significant points that are driving Wolf's account, and to develop principles of responsibility that reflect both the significant aspects of Wolf's account, and the important aspects of the broadly Millian View of moral reasons that I will defend in this chapter.

3.1 The Normative Force of Moral Obligations

3.1.1 The Millian Approach to Moral Normativity

Although the discussion so far has been largely anti-rationalistic when it comes to explaining the force of moral obligations, there are aspects of the Millian-influenced view I want to defend that might seem to fall within the broadly rationalistic tradition. Mill's argues:

> "We do not call anything wrong unless we mean to imply that a person ought to be punished in some way 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".[11]

I mean to take this claim in a weaker sense than the way Mill might have understood it himself. There are certain attitudes that we may think are importantly linked with the idea of moral wrongness, and also with the idea of moral responsibility. The appropriateness of those attitudes seems to be part of what explains the normative pressure associated with moral duties.

The attitudes I have in mind are those that P. F. Strawson calls "reactive attitudes". These are the kinds of attitudes that characterise the overall system of interpersonal demands and expectations involved in human interaction. These include feelings such as resentment, gratitude, remorse, reciprocal love, and hurt feelings. I want to consider what the difference is between the times in which we would consider such attitudes to be appropriate, and the times in which we would not.

Strawson takes reactive attitudes to be fundamental both to our sense of humanity, and to our understanding of moral responsibility and what's required for it. The concepts of moral praise and blame are closely tied up with the kind of reactive attitudes we associate with people's different responses to the morally relevant claims of others. They are generally considered to be reactions to the quality of a person's will towards ourselves and others.

In relation to such attitudes, it seems there is a sense in which certain emotional responses can count as rational or irrational. Although we might think that we cannot accuse someone of irrationality for *failing* to feel an emotion that they *would* have some justification for, we *do* tend to think that we can accuse someone of irrationality when they respond with an emotion that's entirely *unjustified.* So whereas, if an agent simply fails to feel resentful when they have been deliberately and

---

[11] Mill, John Stuart. (1861) "On the Connection Between Justice and Utility" in *Utilitarianism* (Everyman: 1910) p. 50

maliciously wronged, this might not justify a charge of irrationality, we *do* think it's irrational to feel resentment towards someone for something that is not their fault.

The sense in which we might consider Mill to be correct then, is in the suggestion that we do not think that an act is an act of wrongness unless we think that other people, or the agent herself, would not be considered *irrational* for feeling the appropriate negative reactive attitude. This condition would supposedly only be met if was some *justification* for the attitude in question. To this extent, there is a clear link between the concept of rationality and the concept of moral rightness and wrongness.

The main difference between the sense in which rationality relates to our reasons for being moral on Mill's view, and the way in which we might suppose them to relate on more classically rationalistic approaches, is that it involves considering what it would be rational for other people to *feel* in response to our act, and seeing a relation between the fact that an act is wrong and the fact that such feelings *would* be rational, but it does not imply that our own reasons for responding to this fact are explained by our being rational agents.

We can perfectly well imagine that an agent is not disposed to meet their moral duties purely because they do not care about the fact that certain reactive attitudes would be justifiable in response to their actions. We might nonetheless think the very fact that such attitudes *are* justified implies that there is (objectively) a reason to meet that obligation. We can suppose this to be true without suggesting that the agent in question, purely in virtue of their rationality, will be motivated by this fact.

So we can accept the kind of link between rationality and morality being drawn by Mill *without* supposing that all rational agents will be moral in virtue of their rationality alone. This means we do not fall prey to Williams's criticisms and we can consistently acknowledge certain significant points made by Hume that we would otherwise have to reject (I will come back to this point).

We may well suppose that our victim would have a good justification for their feelings of resentment towards us, and even that we ourselves would have some justification for feeling remorse, and none of this implies that we will *care* just in virtue of the fact we are rational. We may not be responsive to the feelings of others or to the validity of their reasons for having such feelings, and whilst this will be a very serious deficiency, this Millian account does not imply that this will necessarily be a deficiency in our *rational* capacities. We can accept Mill's explanation consistent with the possibility that some rational agents might not care about some of the demands of morality.

What we do need to suppose, and what a strongly Humean view would not allow us to suppose, is that there is a sense in which we might relate rationality to certain emotions. There is a sense in which it might be rational to feel resentment only when we have been deliberately wronged, or feel guilt only when we have deliberately (or maybe negligently) caused some harm. It's the fact that on certain occasions particular emotions might be seen to be *justifiable*, *appropriate* or *in order* (whether or not we are inclined, by reason or otherwise, to acknowledge this fact) that is our basis for saying that a moral reasons claim is true. *This* fact will not be falsified by a lack of subjective conditions on the agent's part.

3.1.2 Rationality, Moral Normativity & Reactive Attitudes

It seems that rationality functions similarly to morality in terms of its normative force. We can be justly subject to criticism for acting irrationally, and we tend to think that when we consider matters carefully and correctly we will be led to feel some pressure to form beliefs and intentions that accord with the demands of reason. Likewise, we seem to be justly subject to criticism for acting immorally, and we tend to think when we consider matters carefully, this should lead us to feel some pressure to shape our actions in accordance with the demands of morality.

Part of what appears especially significant about moral reasons, however, especially when we view them in relation to moral responsibility, regards the kind of attitudes that we might have towards others in relation to the way they view the demands of morality. In particular, it's in relation to moral concerns that span from judgements about the quality of people's wills towards us that certain reactive attitudes seem justifiable. Such attitudes seem to be in important ways distinct from the kind of attitudes we have in relation to the way others view the demands of rationality. Whilst there is certainly also some similarity, it seems that the difference between these sets of attitudes shows us something about where we take the force of such sets of demands to lie. Specifically, it seems that such categories of demands are explained by quite different considerations.

The reason I want to consider the importance of reactive attitudes is that I think this highlights some fundamental differences between where we take the normative force of morality to lie, and how we think this differs from the normative pressure associated with the demands of rationality. Specifically, I think we need to look at the situations in which certain reactive attitudes are considered to be *in order*, and to think about what explains this; what explains the fact that in some situations certain attitudes might be described as *apt*, *appropriate*, or *justified*, or at least something to that effect.

I want to argue that the reactive attitudes associated with people's responses to their moral obligations are what explain the force of those obligations. Hume argued (very roughly) that certain traits are morally virtuous on account of the fact that they give rise to love and admiration in others, and that other traits are vices because they give rise to hatred in others. This is far too simple. This might explain why some agents have reasons to be moral in Williams's sense of "reason" (in terms of having the right belief and desire - provided such a person happens to *care* about other people's attitudes), but explaining the *normative* force of morality requires more than this.

This is why I want to suggest that what is required is not just that ignoring a moral obligation will give rise to negative reactive attitudes, and that complying with it will give rise to positive ones. We also need such attitudes to be *justifiable*. It must not only be the case that a person would, as it happens, feel some resentment towards me for maliciously and deliberately causing them injury, but also that they would *quite rightly* feel some resentment. Such a feeling would be justifiable; it would *not* count as irrational.

By examining the kinds of attitudes we have towards an agent who ignores the demands of morality, and considering the difference between the kind of attitudes we take to be appropriate in this sort of case, and the kind of attitudes we consider appropriate in the case of those who ignore the demands of reason, we should come to grasp more clearly the extent to which moral normativity and rational normativity are related, and to gauge some idea of the precise role we take the demands of reason to have in explaining our moral obligations.

When a person ignores the demands of morality, we tend to think that certain attitudes are appropriate. Those attitudes tend to be of the kind we would associate with the notion of moral blame. If a person deliberately harms us we will tend to think we have some justification for our feelings of resentment, and we think that the perpetrator would be justified in feeling some degree of remorse. These facts seem to play a significant role in why we think people have reason not to ignore the demands of morality in the first place.

In contrast, if a person does not respond properly to the demands of rationality, and we are hurt as a result, our attitudes do not tend to be the same. If we suffered an injury because our safety was put at risk due to an error someone made in their calculations, we would only feel resentful if we thought that they were *consciously negligent* in making those calculations, or if they had shown a conscious lack of concern for our safety. If we thought such a person had just made an honest mistake in their reasoning, but had not neglected our interests, we might well think that we had

no justification for feeling resentful, and we might not think that such a person was "blameworthy" in the same respect. We might also think such a person had no rational justification for feeling guilty or remorseful.

It's in this sense Hume notes that the relation between reason and morality seems to be quite limited. Hume argued that morality cannot be grounded in reason on account of the fact that feelings of hatred or love, and the associated judgements of blame or praise, seem appropriate in response to morally relevant actions (concerning the ends an agent selects), but would not be appropriate as responses to the quality of the agent's reasoning skills (which for Hume, includes just their capacity for *a priori* reasoning, and their ability to reason about means to certain ends). We can see his defence of this point in the following passage:

> "A person may be affected with passion, by supposing a pain or pleasure to lie in an object, which has no tendency to produce either of these sensations, or which produces the contrary to what is imagin'd. A person may also take false measures for the attaining of his end, and may retard, by his foolish conduct, instead of forwarding the execution of any project. These false judgments may be thought to affect the passions and actions, which are connected with them, and be said to render them unreasonable, in a figurative and improper way of speaking. But tho' this be acknowledged, 'tis easy to observe, that these errors are so far from being the source of all immorality, that they are commonly very innocent, and draw no manner of guilt upon the person who is so unfortunate as to fall into them. They extend not beyond a mistake of *fact* for which moralists have not generally suppos'd criminal, as being perfectly involuntary. I am more to be lamented than blam'd, if I am mistaken with regard to the influence of objects in producing pain or if I know not the proper means of satisfying my desires. No one can ever regard such errors as a defect in my moral character".[12]

My own reasons for drawing this kind of distinction between moral and rational normativity are the same as Hume's. Moral failings tend to be seen as especially serious in a way that rational failings are not. An agent's willingness to neglect the morally relevant claims of others makes appropriate certain reactive

---

[12] Hume, David. (1740), *A Treatise of Human Nature*, Book 3, part 1, sect. 1, par. 12. (Oxford University Press) pp. 295-296.

attitudes that would be entirely unjustified in response to an agent's willingness to form irrational beliefs, or to go about achieving their goals in an illogical manner.

A person's unwillingness to take the morally relevant claims of others into account would seem to give us some justification for certain attitudes, such as resentment on the part of the person harmed, remorse on the part of the agent, and perhaps condemnation of some kind on the part of others generally. These are the kind of judgements we associate with blame. Such attitudes would be entirely irrational or unjustified in response to another agent's honest failings of rationality. If some agent does not use the best means to achieving some end, or makes some error in her mathematical calculations, we would not consider this to be a serious failing in the same way. A person badly affected might feel *upset*, but will have no grounds for *resentment*. The agent may feel *regret*, but will have no grounds for *remorse*. No one will have reason to feel condemnation, and no charge of blame would be justifiable.

This rules out certain forms of rationalism about morality, whereby an agent is taken to be acting morally only to the extent that she is acting rationally, and where we are forced to say that all rational agents will be moral purely in virtue of their rationality alone. But we can accept Hume's point consistent with drawing the kind of link between morality and rationality that we would associate with Mill. If anything, Mill's account seems to support Hume's insights. We can explain the force of Hume's points precisely in terms of the way that the wrongness of an act relates to the fact that we would be justified in feeling certain emotions in response to such an act.

3.2 Millian Theory, Moral Norms & the True and the Good

3.2.1 Moral Reasons & the Acceptance of Norms

I would like to briefly clarify how what I have just been saying relates to the broader aims of this discussion. I want to argue that claims about morality are only

true or forceful on the condition that there is an objective fact about moral value of the kind that Wolf requires as a condition of responsibility. I think that the existence of such objective values is implied by the fact that certain reactive attitudes could be genuinely considered rational or justifiable. I also take it that it *must* be possible for such attitudes to be justifiable in order for moral obligations to genuinely have the kind of normative force we associate with them.

The account that I am defending involves explaining the normative force of moral obligations in terms of the justifiability or aptness of certain reactive attitudes in response to certain acts, and then explaining how certain attitudes could be considered apt or justified in terms of what is objectively valuable – in Wolf's words, "the true and the good". We might want to understand the claim that there is a true and a good in terms of there actually being a correct system of norms; an objectively valid system of values, telling us to treat certain considerations as weighing in favour of certain acts.

In keeping with Wolf's account, the moral reasons that apply to an agent are not to do merely with the system of norms that the agent actually accepts, but are to do with those that are *objectively* valid – irrespective of whether or not the agent accepts this. What actually makes a certain attitude *justified* will depend on "the true and the good".

On this account, the fact that an agent has a reason to perform some act will be explained by the fact that there is an objectively valid system of norms (a true and a good) that weighs in favour of performing that act. It is only on this condition, that certain reactive attitudes in response to an agent's decision to perform the act could be justified. The normative pressure to respond to this reason is explained by the fact that such attitudes would be rationally justified. However, this kind of normative pressure does not rest on the fact that it is *irrational* for us to ignore our obligations. So we do not have to say that all rational agents who have thought about their actions properly will intend to meet their obligations purely in virtue of being rational agents.

Morality will not depend in this kind of a way on rationality for its normative force, but the kind of relation it does involve will allow for the fact that morality has a parallel normative structure to the kind we associate with rationality: All moral agents who have thought about their actions properly will intend to meet their obligations purely in virtue of being moral. We will not be able to accuse those who do not meet their obligations of being irrational, but we could accuse them of being immoral. In fact, this is the only charge that would actually *justify* the appropriate negative reactive attitudes.

This accords much better with the fact that moral failings are seen as more serious than failings of rationality. It also accords well with the fact that the appropriateness of reactive attitudes relates to judgements about the quality of an agent's will towards us. Because of this, the fact that such attitudes are sometimes justifiable also seems to be more to the point when it comes to considering what is actually *wrong* with an immoral act than judgements relating to an agent's rational capacities could be.

Furthermore, there is a genuine basis for criticising a person on these grounds. The annoyance or hurt we might feel in relation to being treated badly is not so lacking in foundation. We would not have a good basis for the kind of reactive attitudes we associate with people's responses to their moral obligations if there were no difference at all between moral reasons and other reasons, and if both were only important in relation to the agent's rational capacities, since these considerations alone fall short of providing any justification for adopting such reactive attitudes.

3.2.2 Punishment & the Expressing of Feelings

So far, I have only been talking about our justifications for responding to certain actions with particular feelings and attitudes. I have not tried to associate moral reasons with any claims about what justification we might have for *expressing*

those feelings, or for taking the fact that we are justified in feeling resentful to justify treating anyone differently.

These issues go beyond merely considering the appropriateness or rationality of reactive attitudes. Expressing our resentment is likely to be hurtful or damaging in some way to the person we feel resentful towards, so this brings us into the realm of actually *punishing* that person. I take it this would require a much stronger moral justification than the kind that's at issue when we talk about what makes a certain reactive *attitude* rationally justifiable.

Whereas I find it plausible to suppose that if some system of norms entitling us to basic moral consideration were taken to be objectively correct, then this in itself would be enough to justify or make appropriate our feelings of resentment towards someone who has caused us deliberate harm, I take it that actually having a justification for *punishing* that person would require significantly more than this. We would have to think that this would itself be supported by the correct system of norms.

Whether or not this could be the case is largely outside of the scope of this discussion. In contrast to the considerations about when particular attitudes are appropriate, I don't take the issue of when certain forms of punishment are justifiable to be fundamental to understanding the force of moral obligations. The claim that actually punishing a person for their moral wrongdoing is justified is a much stronger claim than we need in order to make sense of moral normativity.

3.2.3 Conclusion

In this chapter, I have tried to establish that a broadly Millian view provides an account of moral reasons and the normative pressure we associate with them that avoids some of the problems we encounter when we try to explain such reasons with reference only to what actions are rational from an agent's perspective. It also seems

especially relevant when it comes to understanding the kind of attitudes we take to be appropriate in response to certain acts, and the way this relates to the notions of blame and praise.

This immediately brings us back to considerations that relate to moral responsibility, and to the issue of why we might think that moral reasons are related to moral responsibility. Strawson's considerations about reactive attitudes and their relation to facts about the quality of an agent's will towards others, in particular, seem to be very significant in relation to the initial question of moral responsibility.

In the following chapter, I want to return to Wolf's account of responsibility, and to reconsider certain aspects of Wolf's account in light of the some of the issues discussed in this chapter. I will then try to readdress the question of how we should formulate principles that best account for responsibility, and to reconsider the issue of whether responsibility and determinism are compatible in light of this revised approach.

4 Reason, Reactive Attitudes & Freedom of the Will

The previous chapter aimed to defend a particular understanding of the normative force of moral reasons; one that's distinct from that which we might associate with other reasons. This approach, influenced by Mill, involved linking moral reasons with reactive attitudes of the kind discussed by Strawson. The justifiability or aptness of particular attitudes was linked to there being some correct system of norms in relation to which certain acts might be considered right or wrong.

These considerations about reactive attitudes, and when such attitudes are rationally justifiable, relate to considerations about the quality of an agent's will. This brings us back to issues that are directly relevant to the discussion that we started out with – the issue of freedom and moral responsibility. I will argue that such considerations might lead us to reconsider certain aspects of Wolf's account, which I have so far been defending.

I will argue that many of the features that make Wolf's account so appealing could be accounted for in terms of actual features of the agent's will, and that we can develop an account structurally very similar to Wolf's, but without having to impose alternate-sequence demands for responsibility in cases of blame, and so an account that would be consistent with the truth of determinism.

4.1 Wolf, Strawson & Moral Responsibility

4.1.1 The Aims of this Discussion

To recap, the first chapter began with looking at Wolf's view that a responsible agent is one who is able to respond to objective reasons. In cases of praiseworthy

action, this condition is met automatically when (in the actual sequence of events) the agent does the right thing for the right reason, as this entails that he is *able* to do the right thing for the right reason. For cases of blameworthy action, however, we have alternate-sequence conditions on Wolf's account. We require the outright possibility of an alternate sequence of events, holding all else constant. The fact that the agent did not do the right thing for the right reason is not enough. We need to know that the agent was *able* to do the right thing for the right reason before we can hold her responsible.

The last two chapters were aimed at supporting the idea that we could have objective reasons. Or at least, that we could have objective moral reasons, and so Wolf's account (or something like it) could, if not explain responsibility more generally, at least help to explain *moral* responsibility. I tried to argue that our responses to certain acts could only be considered to be genuinely *justified* or *appropriate* in relation some objectively correct system of moral norms.

This paints a picture influenced by Mill and Strawson, in which the moral wrongness of an act is closely linked to the justifiability of particular reactive attitudes in response to it. On Strawson's view, such attitudes are related to judgements about the quality of an agent's will towards us. If an agent acts in a way which shows that the quality of his will towards us in not in accordance with a correct system of norms, then this may mean we have some justification for feeling resentful in response to being harmed by that act.

In many ways, this picture accords well with Wolf's account. It explains why we might think there are objective reasons, and also why this idea might play an important role in determining when certain reactive attitudes are justified. Furthermore, such attitudes seem to be an essential part of our ordinary understanding of responsibility, and of what it is about acts that makes us consider them praiseworthy or blameworthy.

At the same time, facts about the quality of an agent's will towards others regard only the actual-sequence of events leading to an agent's act. The truth of such facts seems largely independent of whether the agent was able to do anything else. This account does not require the outright possibility of an alternate sequence of events holding all else constant, even for blameworthy acts. This is where the Strawsonian aspect of the picture seems to depart from Wolf's view.

There were numerous questions raised throughout the last three chapters that have been driving this discussion so far, and at this point it might be worth running through these again, as this will help to keep the bigger picture in perspective.

Firstly, there was the issue of whether or not we could even have reasons that are objective in the sense required for Wolf's account, whereby the truth of such reasons claims does not depend on the agent's subjective motives. Secondly, there was the question of precisely how we should explain the normative force of moral obligations if we cannot appeal to any desires the agent already has, in light of which such courses of action might seem rational from that agent's perspective. In trying to deal with this issue, I appealed to the idea of there being an objectively correct system of moral norms, and looked at how this might relate to the justifiability of particular attitudes as responses to certain acts.

These aspects of the discussion are directly relevant to the original issue of responsibility. We now have a whole new set of considerations when it comes to accounting for responsibility, and in light of these considerations, we are in a better position to formulate new conditions of responsibility that account for these factors as well as those features that give Wolf's account its appeal. We can then readdress the question of whether responsibility requires actual or alternate sequence conditions (and hence whether responsibility is compatible with determinism).

In order to answer this question, we will need to look at the way in which Strawson's considerations about the quality of an agent's will, and Wolf's considerations about ability to respond to objective reasons, relate to one another.

4.1.2 Reactive Attitudes, Quality of the Will & Responsibility

On both Wolf's account, and on the Millian/Strawsonian account I have tried to defend, the idea of there being a system of norms that is objectively correct plays an important role. The reasons for this are slightly different on each account, but equally important on both. On the one hand, we need some correct system of norms in order for certain reactive attitudes to be justifiable. On the other hand, an agent needs more than just the ability to respond to their own subjective system of values in order to be held morally responsible for their actions; they need to be able to base their subjective values on the true and the good.

On a broadly Strawsonian account, we respond to an agent with particular reactive attitudes in cases where we have reason to think the agent's actions provide significant information about the quality of that agent's will towards us. For a very young child or someone suffering psychosis, we might think that we cannot make this kind of an inference between their actions and facts about the quality of their will. They are not able to meaningfully understand or engage in the kind of interpersonal exchange within which such attitudes make sense.

This kind of consideration relates to something Hume pointed out: We only consider an act to be blameworthy if it's likely to tell us about the agent's motivations and character. If someone acts through coercion, ignorance, or mental illness, we cannot make this kind of inference. This fits well with the idea that an action has to tell us something about the quality of an agent's will towards others if it's going to be rational for us to respond with any particular reactive attitude.

In relation to Wolf, it seems that similar considerations play a role in some of her examples. She notes that we would not hold an agent responsible for pressing a button that would electrocute someone in the next room if we thought that he did not know that this is what pressing the button would do. Likewise, we would not hold someone responsible who did not know that being electrocuted was a painful or unpleasant experience. But further to this, we would not hold an agent responsible who (as difficult as this is to imagine) could not understand that it was *wrong* to cause pain to people.

If an agent is led, through her unimaginably horrendous upbringing, to have absolutely no concept of why we should show any kind of consideration or respect towards others, she may end up accepting a very skewed system of norms. Her values will be determined by her disturbing upbringing, and not by the true and the good. Because of this, whether or not she is able to act in accordance with the system of norms she actually accepts will be irrelevant to whether we should hold her responsible. We need to know whether she is able to act in accordance with the system of norms that's actually valid – with the true and the good. If she doesn't *know* that causing pain is wrong, then supposedly she cannot reasonably be expected to shape her motivations in accordance with this fact.

So we might think that a genuinely good-willed agent is one who is motivated to *do the right thing*, whatever that may be. A good-willed agent will want to treat others the way that he morally ought to. But a person who, through no fault of his own, has been led to accept a system of values that differs drastically from the values there actually are, will not really have the opportunity to act in accordance with the true and the good, even if he wants to. In such a case, his willingness to cause pain might not tell us much about whether he is motivated to do the right thing. If his values are momentously flawed, we will not know whether he is good-willed enough that he would have done the right thing had he actually known what the right thing to do was. It's only if he's actually *able* to do the right thing that it is fair to hold him responsible when he fails to.

It seems then, that the issue of how *good-willed* the agent is plays an important role in both the kind of considerations that are motivating Strawson's account, and the kind that are motivating Wolf's.

For Strawson, the fact that an agent is able to meaningfully engage in interpersonal exchange, and to properly understand the kind of demands that characterise people's moral expectations of one another, is enough to tell us that when that agent neglects the morally relevant claims of others, he does so despite the fact that he understands the moral significance of his actions. This gives us reason to think that such an agent is not good-willed. Negative reactive attitudes are a direct response to this.

On Wolf's account, if an agent is able to act in accordance with the true and the good, this tells us that if she chooses *not* to do so, this is genuinely an expression of her *unwillingness* to do so. This also tells us such an agent is not good-willed.

However, on a Strawson-style account the test is whether the agent understands the moral significance of what he is doing, whereas on a Wolf-style account, the test is whether the agent is capable of doing anything better. (If an agent does not understand the significance of what he's doing, this might be important, but just *because* it implies the agent is not capable of doing anything better).

Both accounts seem to represent perfectly intelligible rationales for judging something that seems to be significant in determining when it is rational to feel resentment in response to some agent's act – namely, whether the act is really an expression of bad-will on the agent's part. However, Wolf's condition is an alternate-sequence one, and Strawson's condition is an actual-sequence one.

4.2 Considering What Matters Most for Responsibility

4.2.1 Conditions of Responsibility

Whatever precise conditions of responsibility are the important ones, the agent's ability to *know* right from wrong seems to be a key feature. The reason it seems that we do not want to hold an agent responsible who is incapable of engaging in the kind of interpersonal exchange that characterises our moral expectations of one another, is that such an agent cannot possibly *know* what the moral significance of their actions are. Likewise, on Wolf's account, whether the agent is able to respond to the reasons that there are is very closely linked with whether or not that agent *knows* what reasons there actually are.

On Strawson's account, we might think that the condition of being able to engage in the overall web of reactive attitudes is important *because* it tells us about whether or not that agent can be expected to fully comprehend what constitutes right and wrong. If so, knowing about moral reasons will be the more fundamental consideration for judging what some act tells us about the quality of an agent's will. However, on Wolf's account, the key factor is the agent's ability to respond to objective reasons. Whether or not the agent understands the difference between right and wrong is only important insofar as it tells us something about her ability to respond to the reasons there are.

The kind of examples Wolf uses to demonstrate her point often involve agents who have a misguided set of values, and so do *not* actually know what reasons there really are. Wolf takes this to be significant because of the fact that such an agent could not reasonably be expected to be capable of responding to objective reasons. But we might wonder whether this really is the important point, or whether it's merely the fact that the agent doesn't have a good grasp of right and wrong to begin with that accounts for the intuition that such an agent cannot be held responsible for his actions.

If it's just whether or not the agent knows the difference between right and wrong that matters, we will only require an actual-sequence condition for responsibility. If this only matters insofar as it tells us about the agent's ability to do otherwise, however, we will also require the outright possibility of an alternate sequence with certain features. But we now have a task similar to the one that concerned Hurley. We have an actual-sequence principle and an alternate-sequence principle, and we need some way to work out which one is actually doing the work in determining whether or not an agent is responsible.

Hurley tried to work out whether it was Fischer's condition of acting on a reasons-responsive mechanism or Wolf's condition of being able to respond to objective reasons that matters. This case is somewhat different. We have Wolf's principle, and one that's similar to it, but which does not impose any alternate sequence demands. We want to know whether it's the fact that an agent knows what reasons there are or the fact that an agent is able to respond to those reasons that determines whether she is responsible.

However, this might be problematic. The fact that an agent is able to respond to objective reasons seems to imply that that agent knows what reasons there are. Likewise, if an agent does not know what reasons there are, it seems that we could not reasonably expect her to respond to those reasons.

The other way round, we do not have this problem, but we run into different ones. An agent may well know what the right thing to do is, but be unable to respond to such facts because of coercion or restraint of some kind. But clearly, this in itself will not show that it's only ability that matters. If knowledge of right and wrong is important in terms of what it tells us about the quality of an agent's will, then we would need the agent's response to those reasons to be a genuine response to their understanding of right and wrong. If the agent is not in a position to be acting *in light of* their state of knowledge, then such actions will not be relevant.

Likewise, if an agent is restrained from doing the right thing, this doesn't imply that he won't know what the right thing to do is, but it does mean that his actions will not qualify as being those actions he chooses *despite* his knowledge of what's right. They will instead be actions he is forced into completely irrespective of such knowledge.

Supposing some agent has made a promise to meet someone who has helped him greatly, and to whom the agent knows this meeting is very important. He may be fully aware of having an obligation to keep his promise to meet her at the arranged time, and may fully understand the significance of his failing to do this, fully grasping how hurt her feelings will be. But if, through no fault of his own, the agent is locked in a room with no way to escape and no way to contact the person to whom this duty is owed, we would not hold him responsible for his failure to turn up. If the person he was meeting found out about his predicament, we might expect her to feel disappointment, but hurt feelings or resentment would seem out of place. In this case, we can see that the agent's grasp of right and wrong alone is not enough.

However, we don't need to say that such an agent could only be responsible on the condition that he had been *able* to do the right thing for the right reason either. It might still be that the agent's understanding of right and wrong plays an important role, but it plays this role only insofar as it tells us something about that agent's *willingness* to do the right thing – and willingness to do the right thing seems clearly distinct from ability to do the right thing. We might think that the act is not one the agent is responsible for because it doesn't tell us anything about the kind of value that agent places on doing what he knows is right.

In this kind of case, it seems the important factor is the agent's level of willingness to do the right thing. If we knew that such an agent was very willing to do the right thing, and that he placed a very high value on keeping his promises, showing gratitude where it's owed, and not hurting people's feelings, we would have reason to suspect that had he been able to do the right thing, he would have done. But we

might think that it's neither the ability to do the right thing, nor the truth of the counterfactual "he would have acted rightly, if he could have done" that matters.

These features might only be important because they tell us something about the value the agent places on doing what he knows to be right. If the agent is restrained beyond his own control from keeping his promise, then his failing to meet this duty tells us nothing about his attitude towards what he knows to be right and wrong. His act is irrelevant to what he understands to be the right thing to do. So perhaps we would only hold him responsible if we thought he was consciously acting *despite* the fact that he knows what he does is wrong.

It seems there is a clear difference between an act that happens *irrespective* of an agent's knowledge, and an act that the agent performs *despite* his knowledge. In the former case, the agent's reasons for acting are such that they do not relate to the value that agent places on his understanding of right and wrong. In the latter case, the act results precisely from unwillingness on the agent's part to accord the kind of value to a course of action that he knows it to be worthy of.

On any account, we might have good reason to suppose that knowledge of right and wrong is a necessary condition of responsibility, but just stated in these simple terms, it's not enough to count as a sufficient condition. We need more than just the fact that the agent knows right from wrong. As the above example shows, if she is going to be held responsible for her act, we need that act to count as an expression of her attitude towards that knowledge. We need it to relate to her *willingness* to act in accordance with what she knows is morally correct.

I want to suggest that we reconstruct the actual-sequence principle as follows: The agent will be responsible for an act if she has a good understanding of what reasons there are objectively *and* that act is genuinely an expression of her actual willingness (or unwillingness) to take those reasons into consideration.

4.2.2 Comparing Conditions

We now need to work out which condition is closer to what actually accounts for responsibility; whether it's the fact that an agent knows right from wrong, and his act is an expression of his level of willingness to take this into account, or rather the fact that he is able to respond to the reasons there are.

The fact both seem to be relevant considerations may be best explained by supposing that one of these is important in itself, and one is important simply due to its relation to the other. But we do not yet know which should be considered more fundamental. Is the fact that an agent is unable to respond to objective reasons only important because it tells us that their action is not an expression of their level of willingness to place the correct value on their understanding of right and wrong? Or is it that their act being an expression of their willingness to place the correct value of their understanding of right and wrong is only important because it tells us that they are capable of responding to objective reasons?

It seems it could only be the actual-sequence condition that is more fundamental, and the alternate-sequence condition that is only important because of what it tells us about whether the agent meets the actual-sequence one. The agent's *ability* to act in accordance with the reasons there are is important *because* it implies that the agent's failure to do so is an expression of their attitude towards those reasons; an expression of the importance they place on treating others the way that they know they morally ought to.

I think that it must be this way around because the inference would not be valid the other way around. We cannot infer from the fact that an agent's act is an expression of their unwillingness to take considerations of right and wrong into account that such an agent is capable of making a better decision, and yet this still seems to be a relevant consideration.

If we suppose the thesis of determinism to be true, this would entail that there were never cases in which an agent is capable of doing something other than what she does in fact do. This would mean that if an agent does not do the right thing for the right reason, she is not able to do the right thing for the right reason, since it will never be an outright possibility holding all else constant. But this does not entail that such an agent doesn't act *despite* her understanding of right and wrong. It might still be that her action is an expression of her unwillingness to treat certain possible courses of action in a way that reflects the moral value she knows them to be worthy of.

The relevance of this condition cannot be explained by its relation to considerations about an agent's *ability* to do the right thing for the right reason. It seems to be important in its own right. This is hardly a conclusive argument to show that ability is not important. But it does rule out the idea that the actual-sequence principle is only important because of what it tells us about the agent's ability to respond to objective reasons. This leaves us with some pretty good reasons to favour the actual-sequence condition.

Firstly, if someone were to find conclusive evidence that our actions were causally determined (and hence Wolf's alternate-sequence condition was never met), it seems implausible to suppose that the actual-sequence considerations about whether an agent fully grasps right and wrong, and whether their actions are expressions of their willingness to take this into account, would become entirely *irrelevant* to the question of moral responsibility. We would at least still think that such considerations played an important role in determining when it would be rational for the victims of certain acts to feel resentful in response, or for the perpetrators to feel remorseful.

Secondly, we can account for our intuitions about which cases we should consider an agent to be morally responsible in, purely with reference to facts about that agent's grasp of right and wrong, and whether the act is an expression of her willingness to take this into account, without *having* to commit ourselves to indeterminism.

If we *can* adequately differentiate between cases where we do not want to hold an agent responsible and cases where we do, just by looking at features of the actual sequence, we might think this undermines our largest motive for adopting a condition that requires the outright possibility of an alternate sequence. If we don't have to make this strong a claim in order to sustain an account of responsibility that works, this seems to be a good motive for rejecting it. It's unnecessary to hang onto an account that's consistent with less metaphysical possibilities despite the fact it has no *extra* explanatory force that we cannot find without it.

4.2.3 A Reformulated Account

If we accept that considerations about the quality of an agent's will are what matters most for moral responsibility, and that this is best assessed with reference just to an agent's state of knowledge with regard to right and wrong, and with whether or not the agent's act is an expression of his willingness to take this knowledge into account when deciding how to treat others, we end up with an account somewhat different both to Wolf's account, and to the kind of account that Strawson offers. I will try to summarise exactly what the distinctive features of this view of responsibility are.

On this view, we distinguish relevant features of the agent and relevant features of the act. In order for an agent to be responsible for an act, we have conditions that relate to both, each of which is necessary for responsibility, and the conjunction of which is sufficient for it. It's necessary that the *agent* has a good grasp of what reasons there are. It's also necessary that the *act* is one that's actually an expression of her willingness to take into account her knowledge of what's morally right in deciding how to treat others. If both of these conditions are met, the agent will be responsible for the act.

Both of these conditions are asymmetric in a very similar way to Wolf's condition, but only if we take into account the importance of acting on the right *reason*. I will try to demonstrate what I mean by this in relation to both the agent-centred and the act-centred conditions.

Wolf argues that an agent can only be held responsible who is able to do the right thing for the right reason. On this agent-centred condition, in contrast, an agent can only be held responsible who *knows what it is* to do the right thing for the right reason. For Wolf, if an agent does the right thing for the right reason, her condition has automatically been met, since actually doing the right thing for the right reason automatically implies that he is *able* to do the right thing for the right reason.

On this account, the condition is not met merely because the agent does the right thing, but as with Wolf's account, it seems that it is met when an agent does the right thing *for the right reason*, since this implies that the act is based on her understanding of what is right. Supposedly, the *right* kind of reason to do something will have to relate in some way to the fact that the agent knows it's the right thing to do.

An agent might do the right thing because she is acting *despite* her flawed system of norms, and actually doesn't realise she is doing the right thing at all. E.g. she has been hypnotised to believe that the polite thing to do when meeting someone for the first time is to punch them in the face, but she chooses not to, despite thinking that it's the right thing to do, and instead shakes their hand. She does this merely because she considers herself to be a bit of a rebel and wants to cause offence. But supposedly in this kind of case, she will not be acting for the right kind of *reason* – she will not be acting on the basis of her knowledge of the true and the good. So for both Wolf's condition, and this account's agent-centred condition, if an agent does the right thing specifically *for the right reason*, we have already established she is responsible for her act (although on this account merely *doing* the right thing is not enough if we do not know what kind of reason she was acting on).

With regard to whether the agent's *act* is really an expression of the importance that agent places on treating others in the way she ought to (the act-centred condition), it also seems like this condition is automatically met in the case of praiseworthy action, where the agent does the right thing specifically *for the right reason.* Again, this is because the agent would *not* be acting for the right kind of a reason if that agent's act were entirely *irrelevant* to her understanding of the true and the good. E.g. the agent keeps her promise to stay indoors today, but only does so because she has accidentally locked herself in the house.

Since supposedly, the agent will only be acting for the right kind of a reason on the condition that her understanding of the true and the good is actually playing some role in motivating her to perform the act, the fact she has done the right thing *for the right reason* already implies that her act is an expression of the importance she places on doing the right thing.

So this account is similar to Wolf's in that (so long as we take it to be important in praiseworthy acts that the agent does the right thing for the right reason) it preserves the asymmetry between cases of praise and blame. For cases where the agent fails to do the right thing for the right reason, the conditions of responsibility will not be met automatically. In order to know whether an agent is responsible for his act, we need to know whether he has a good grasp of right and wrong, and whether his act is an expression of the value he accords this.

This account also preserves the importance that Wolf places on there being a true and a good, and on this playing some role in the agent's reasons for action. It holds on to the supposition that there is some correct system of norms, and that the agent needs to bear the right kind of relation to this correct system of norms in order to be held responsible for his act. The difference is just to do with the precise relation that's taken to be relevant. For Wolf, the agent must have the *ability* to shape her actions in accordance with the true and the good, whereas on this account, the agent

must have a good understanding of the true and the good, and her actions must express her *attitude* towards what she knows to be morally right. But she does not need the outright *ability* to act in accordance with the true and the good, keeping all else constant.

This is where this account also bears a close relation to Strawson's. On Strawson's account, considerations about whether or not an agent is able to do otherwise are largely irrelevant to anything that we take to be important to us about moral responsibility. The considerations that seem to matter are those that concern the quality of the agent's will towards others. For judgements about the quality of an agent's will to matter or make sense, that agent needs to be able to meaningfully engage in the exchange of reactive attitudes that characterise our moral demands and expectations of one another.

This seems to be relevant not *only* because such agents cannot respond in any constructive way to the kind of reactive attitudes that would be relevant to their actions, but *also* because such actions do not tell us much about the quality of an agent's will. If an agent cannot understand this kind of interaction, he will not be able to have an adequate grasp of the moral significance of his actions, and so will not have a good understanding of how he *ought* to treat others. We cannot infer from such an agent's actions anything about whether that agent is concerned to treat others in the way that he ought to. He is not responsible, because he cannot meaningfully engage with the kind of considerations that are relevant to understanding what he ought morally to do.

This strategy of thinking about the agent's knowledge of right and wrong, her attitudes when it comes to taking such considerations into account, and what this tells us about the quality of her will, focuses responsibility on the actual sequence. This is why such an account, although in various ways very similar to Wolf's, is a compatibilist account. We do not require the outright possibility of an alternate sequence of events, even for cases of blameworthy action.

4.3 Frankfurt's Case & Hurley's Question

4.3.1 Reconsidering Hurley, Fischer & the Frankfurt Example

At this point, it might be worth saying a bit about Frankfurt's case. The problem I outlined in the first chapter, with using the example to show that even in the case of determinism, the existence of alternate possibilities could be seen as irrelevant, is that determinism poses distinct problems that cannot be accounted for so easily. In the case of determinism, we are concerned about the possibility of some alternate sequence, whereas in the Frankfurt case, we are looking just at its content.

The fact that the cases differ in this way, leaves open that some alternate-sequence principle (such as that offered by Wolf) could be relevant in a way that the example does not challenge. Unlike with Frankfurt's case, determinism does not just concern what *would* happen if an agent chooses not to perform some act, it also affects whether or not that agent actually *does* perform that act. And unlike Frankfurt's counterfactual intervener, we cannot just add or subtract determinism from our example and assume that an agent's actions can be assessed in the same way in either case.

Hurley's strategy to show that Frankfurt's irrelevant alternative intuition *does* generalise to the case of determinism, involved arguing that whereas the agent's meeting the actual-sequence condition makes a considerable difference to whether we want to hold the agent responsible for her actions, whether or not the agent meets the alternate-sequence principle makes no difference at all. But because her alternate-sequence principle involved considering what the agent *does* in a possible world where determinism is false, her account rested on the kind of answer she thought we could give to a question that simply cannot *be* answered.

Hurley's mistake was in thinking that such a question *could* be answered, and this is because she thought such an answer could be inferred from features of the actual-sequence. But this inference appears to be faulty, and that is because it turns out to be an instance of the conditional fallacy. She tries to work out what the agent would do if she could have acted on an entirely different mechanism by looking at the features of the actual mechanism, when these might play no role at all in the agent's choice in the alternate sequence. If indeterminism *was* true, we could no longer rely on features of the *actual* mechanism being present, as the truth of such a thesis may well make it false that the agent acts on this mechanism in the first place.

We might, however, think that the account of responsibility I am defending explains why Hurley's question seems compelling in the first place. If we knew that an agent would do exactly the same thing even if he could choose not to, this might imply that the act is genuinely an expression of the importance that he places on acting as he does. If he has a good grasp of right and wrong, then, the act would genuinely be an expression of his willingness to take this into account.

Whether the agent knows right from wrong to begin with is not considered a necessary condition on Hurley's account. Instead, she favours Fischer's condition. The agent needs to be acting on a sufficiently reasons-responsive mechanism. So for Hurley, if an agent acts on a sufficiently reasons-responsive mechanism, *and* that agent would act on exactly the *same* mechanism in a nearby possible world where he could have acted on an alternate one, then we can hold him responsible for his act.

Hurley's account is in some respects similar to the one I am defending. However, both of her conditions suffer serious problems that could be avoided. On the one hand, her first condition, is one for which there may be no answer as to whether or not it is fulfilled, since we simply have no way to settle the question of whether the agent would do otherwise where this entails acting on a different mechanism altogether. On the other hand, her Fischer-influenced second condition of acting on a

reasons-responsive mechanism is analogous to Nozick's condition for knowledge, and so is subject to the same kind of counterexamples that affect Nozick's account.

Now we have developed new conditions of responsibility, we should be able to generate analogous counterexamples by considering cases in which these conditions and Fischer's conditions come apart: cases where the agent acts on a reasons-responsive mechanism, but the agent's act is not a reflection of their attitude towards what they know to be morally right, and vice-versa.

To recap, Fischer's reasons-responsiveness conditions are as follows:

(1) In other (on maximal responsiveness, nearby) possible worlds where there is reason to perform the act, the same mechanism would lead the agent to perform that act.
(2) If it were not the case (in nearby possible worlds where it's not the case) that there was reason to perform the act, the agent would not have performed the act.

We can see that these conditions cannot be necessary or sufficient for responsibility when we consider the following counterexamples.

Firstly, an agent may be responsible despite failing condition (1).

A man donates money to a worthy charity out of a sense of social duty. There is optimal reason for him to do this. However, there have been false reports published, claiming that a particular company will double any donations made to that charity if you give the money to them. Actually they are just stealing money that would go towards a worthy charity. Someone distracts this man just before he gets the chance to read these reports. Had he read them, however, he would have believed them, and so if he had acted on his sense of social duty, he would have given his money to this corrupt company instead of giving it to charity.

In this case, he would fail condition (1). Yet it still seems that we should hold him responsible for his act when he *is* doing the right thing. He acts in light of his understanding of why he has reason to give his money to the charity, on the basis of his sense of social duty. The fact that this mechanism (his sense of social duty) would lead him to do something else in a nearby possible world should not diminish his responsibility in any way. The act should still count as praiseworthy, as it reflects his willingness to do what he knows he morally ought to.

It also seems that an agent may be responsible despite failing condition (2):

An agent owes her friend £50. She gives her friend the money on the basis of being overly anxious to always promptly pay back any money she owes. As it happens, she is so overly anxious in this respect, that the same mechanism of being anxious to pay back money promptly, would have led her to persuade herself that she owed this person £50, and to give her the money, even if she had not owed her any money, and had no good reason to perform such an act. Despite this, it seems that we would have good reason to hold her responsible for her act. Her act would still be just as praiseworthy, as it expresses her willingness to do what she knows she morally ought to.

It also seems that an agent could pass both of Fischer's conditions, and still not be responsible for her act.

Lucy is a paranoid schizophrenic who has a tendency to form wildly unlikely paranoid delusions about her neighbours trying to kill her on the basis of insufficient pieces of evidence. As a result, all of those in her community are very careful to make sure that she never experiences anything at all suspicious or out-of-the-ordinary, so as not to set her off. One day, she looks out of her window, and she sees someone on the street stop for a moment and glance in her direction (the kind of thing people around her would generally avoid doing in case they trigger any more delusions). She

concludes from this event that her next door neighbour is part of an elaborate conspiracy to kill her, and that he has paid this person to check that she is at home so that he can come round and shoot her. She decides she had better quickly go to his house and kill him, in self-defence, before he gets the chance to kill her first. As it happens, she was absolutely right, and all of these beliefs were true.

If it hadn't been the case that her neighbour wanted to kill her, and that she had reason to kill him first, she would not have seen someone glance in her direction (since, unbeknownst to her, those in her community go to enormous efforts not to do that kind of thing), and so she would not have been subject to the delusion, and she would not have killed her neighbour. Any other situation where she had reason to kill her neighbour, and where she was acting on the basis of the same mechanism – her paranoid schizophrenic delusions about her neighbour's intention to kill her – would have led her to do the same thing. So she passes both of Fischer's conditions. But it would not be right to hold her responsible for her act because she was acting on the basis of her paranoid schizophrenic delusion. The fact that such a mechanism, in her situation, happens to track reason doesn't make her responsible for her act.

In these examples, the conditions under which the mechanism is responsive to reasons come apart from the conditions under which the agent's act is based on their willingness to take into consideration their understanding of the true moral significance of the act. Here we find that it's the latter conditions that determine when we should hold the agent responsible. So it seems the Fischer-influenced condition Hurley adopts is only relevant in the cases where it might imply that the agent understands what is right and that their act is an expression of their willingness to take this into account.

It seems then, both of Hurley's conditions are really aimed at trying to determine something more fundamental – whether the agent is basing their actions on the value they place on what they know to be morally correct. Knowing that an agent *would* not do the right thing even if she *could* only seems relevant if we think it tells us something about that agent's willingness to treat others as they know they morally

ought to. On it's own in relation to determinism, it's not even clear that it's a coherently answerable question, so it's force must relate to what it tells us about something that we *can* give an intelligible answer to.

4.3.2 Responsibility for Skewed Values

The idea that an agent will not be responsible if she is acting on the basis of a faulty system of values might seem problematic. Sometimes it seems like our resentment towards a person is justified precisely on the basis of the fact that she adopts a skewed system of values. E.g. a guard at a Nazi concentration camp might well fully believe in what she is doing, and believe that her actions are really in accordance with the true and the good. Just the mere fact that they are *not* in accordance the true and the good, on its own, certainly shouldn't be enough to get her off the hook. We might feel even more contempt for her because of the fact that she could actually accept this system of norms in the first place.

The issue here will have to turn on whether or not we actually hold the agent responsible for her acceptance of a particular set of norms to begin with. We will need to ask whether she is responsible for her own ignorance regarding what is actually true and good.

The adopting of a particular set of norms will have to be considered an action assessable in the same way as any other. We will need to know whether she understood her duty to develop a good understanding of certain kinds of relevant information when she formulated her current system of values, and whether her adopting the norms she does is based on her failure to take this duty into consideration.

The case of becoming a Nazi seems especially striking because it seems as though the relevant information about the wrongness of the act is quite readily

available. Anyone whose upbringing has allows them a basic grasp of human rights at all seems to be equally in a position to apply that information to another race that share the same morally relevant features. This is why extreme racism seems especially difficult to forgive. It seems like anyone who can apply the correct level of moral consideration to their own race, and has basic reasoning skills, should have all the relevant information available to them in order to grasp that the same level of moral concern is owed to others. We cannot imagine that there were many Nazis who lacked the basic understanding of right and wrong required to formulate better ideals to begin with.

It's worth noting, however, that we can *imagine* cases where these conditions for responsibility were not met. E.g. we might not hold her responsible if we found out that from a very young age she was taught that all other races were actually robots programmed to *act* just like humans, but who actually did not have any more conscious experience than your average toaster (we must suppose that she had no access to arguments for computational theories of the mind that might lead her to suspect that even robots could have conscious experience). If this information was perpetuated in an elaborate conspiracy throughout her entire life, and she had never heard anything that could lead her to suspect this was false, then we might think she was not responsible for adopting a system of norms that gave her own race a superior moral status.

Likewise, if we thought her adopting of skewed norms was not an expression of the importance she placed on formulating a system of values that was based what she knew to be morally right, we would not hold her responsible. E.g. if she was brainwashed or hypnotised into accepting her skewed system of norms.

Wolf often uses the example of an agent who had never once in his life witnessed anyone showing the slightest bit of kindness or consideration to anyone else. She also supposes that whenever he showed any kindness himself, this was taken

advantage of and he suffered as a result. We might understand if this person thought that the only valuable course of action involved being entirely selfish.

This would be a very extreme case. It's hard to imagine anyone actually being this deprived of moral input. But if an agent was actually brought up in such a situation, we clearly could not blame him for lacking any grasp of what is morally right and wrong. If he were to cause pain to another person, we would have to consider this a terrible consequence of the horrendous situation he was led to formulate his system of values in. We could not hold him responsible for his act.

In real life, we do not expect that people are dealing with such an extreme lack of moral input, but nonetheless, there will clearly be some cases in which an agent cannot be expected to encounter the information available to base *some* of their norms on what is actually true and good. In such cases, it would not be fair to hold the agent responsible for the acts that she performs on the basis of those skewed values. However, there may also many cases where we *do* think agents are responsible for their own ignorance about right and wrong, and in such cases the mere fact they have a skewed system of values alone will not get them off the hook.

We will need to add a clause to our conditions. If a condition seems to have been failed, we also need to know whether the agent is responsible for this fact. This will be determined by applying these same conditions to the decisions that led the agent to his current values and attitudes.

4.3.3 Conclusion

It seems we have good grounds to suppose that the factors that really matter in attributing responsibility are firstly, that the agent has a good understanding of what is morally right and wrong, and secondly, that the act is genuinely an expression of the importance that agent places on treating others in the way she knows that she morally

ought to. If the agent fails either of these conditions, he will not be held responsible, unless of course, he is responsible for his own failure of the conditions (and this is something we can judge by the same criteria).

This differs both from Strawson's view in some respects, and from Wolf's view in some respects, but is aimed at assessing something that seems to be fundamental on both accounts: what the act tells us about the quality of an agent's will towards others. Specifically, what it tells us about that agent's level of willingness to treat other people the way they morally ought to.

This also seems to account for what is driving Hurley to ask about what the agent would have done had that agent been able to do otherwise. But this question is not one that we can give a meaningful answer to in relation to determinism, and so the account I am defending avoids this problem, and also accounts for responsibility in cases where Fischer and Hurley's reasons-responsiveness principle fails.

Bibliography

Fischer, John Martin. (1987) "Responsiveness and Moral Responsibility", in *Responsibility, Character and the Emotions*, ed. by Ferdinand David Shoeman (Cambridge University Press) pp. 81-106.

Fischer, John Martin, and Ravizza, Mark. (1998) "Moral Responsibility for Actions: Weak Reasons-Responsiveness" in *Responsibility and Control*, chapter 2 (Cambridge University Press) pp. 28-61.

Fischer, John Martin, and Mark Ravizza. (1998) "Responsibility for Actions: Moderate Reasons-Responsiveness" in *Responsibility and Control*, chapter 3 (Cambridge University Press) pp. 62-91.

Frankfurt, Harry. (1969) "Alternate Possibilities and Moral Responsibility", *the Journal of Philosophy*, vol. 66, pp. 828-39.

Frankfurt, Harry. (1997) "Equality and Respect" in *Necessity, Volition, and Love*, chapter 13 (Cambridge University Press: 1999), pp. 146-154. Especially Sect. 3, pp. 152-154.

Hume, David. (1740), "Moral distinctions not deriv'd from reason" in *A Treatise of Human Nature*, Book 3, part 1, sect. 1, (Oxford University Press) pp. 293-302. Especially par. 12, pp. 295-296.

Hume, David. (1740), "Of the origins of the natural virtues and vices" in *A Treatise of Human Nature*, Book 3, part 3, sect. 1, (Oxford University Press) pp. 367-378.

Hurley, Susan. (2000) "Reason, Responsibility and Irrelevant Alternatives", *Philosophy and Public Affairs*, vol. 28, pp. 205-24.

Mill, John Stuart. (1861) "Of the Connection between Justice and Utility" in *Utilitarianism*, Chapter 5, in the collection *Utilitarianism, On Liberty, Considerations on Representative Government*, ed. by Geraint Williams (Everyman: 1910) pp. 43-67.

Moore, G. E. (1912) "Free Will" in *Ethics*, chapter 6 (Oxford University Press: 1965) pp. 84-95.

Nozick, Robert. (1981) "Knowledge and Scepticism" in *Epistemology: An Anthology*, ed. by Ernest Sosa and Jaegwon Kim (Blackwell: 2000), pp. 79-101, especially the section on knowledge, pp. 79-86.

Raz, Joseph. (2002) "On Frankfurt's Explanation of Respect for People",' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. by S. Buss & L. Overton, (MIT Press) pp. 299-315.

Scanlon, Thomas. (1998) "Appendix. Williams on Internal and External Reasons" in *What we Owe to Each Other* (The Belknap Press of Harvard University Press) pp. 363-373.

Scanlon, Thomas. (1998) "Reasons" in *What we Owe to Each Other,* chapter 1 (The Belknap Press of Harvard University Press) pp. 17-77, especially the introduction, pp. 17-18.

Scanlon, Thomas. (1998), "Responsibility" in *What we Owe to Each Other,* chapter 6 (The Belknap Press of Harvard University Press) pp. 248-244, especially the introduction, pp. 248-241.

Shope, Robert K. (1978) "The Conditional Fallacy in Contemporary Philosophy", *The Journal of Philosophy*, vol. LXXV, Number 8, pp. 397-413.

Strawson, P. F. (1962) "Freedom and Resentment", *Proceedings of the British Academy*, vol. 48, pp. 1-25.

Williams, Bernard. (1981) "Internal and External reasons" in *Moral Luck,* chapter 8 (Cambridge University Press) pp. 101-113.

Wolf, Susan. (1980) "Asymmetrical Freedom", *the Journal of Philosophy*, vol. 77, pp. 151-66.

Wolf, Susan. (1990) "The Real Self View" in *Freedom Within Reason*, chapter 2 (Oxford University Press) pp. 23-45.

Wolf, Susan. (1990) "The Autonomy View" in *Freedom Within Reason*, chapter 3 (Oxford University Press) pp. 46-66.

Wolf, Susan. (1990) "The Reason View" in *Freedom Within Reason*, chapter 4 (Oxford University Press) pp. 67-94.

Wolf, Susan. (1990) "The True and the Good" in *Freedom Within Reason*, chapter 6 (Oxford University Press) pp. 117-147.