Moral Responsibility and the Self

by

Thomas Blanchard

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Arts
in
Philosophy

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

Moral responsibility is an issue at the heart of the free-will debate. The question of how we can have moral responsibility in a deterministic world is an interesting and puzzling one. Compatibilists arguments have left open the possibility that the ability to do otherwise is not required for moral responsibility. The challenge, then, is to come up with what our attributions of moral responsibility are tracking. To do this, criteria which can adequately differentiate cases in which the agent is responsible from cases in which the agent is not responsible are required. I argue that an agent is responsible for the consequences of an action if they stem, in an appropriate way, from the agent's deep values and desires. These deep values and desires make up the Deep Self. Parts of the Deep Self, first, tend to be enduring; second, desires within it tend to be general (as opposed to directed towards specific things); third, they tend to be reflectively endorsed by the agent; fourth, these traits are often central to the agent's self-conception; and fifth, they are not generally in extreme conflict with other deep traits. Empirical work is drawn upon to help develop a suitable account of what deserves to be called a part of the Deep Self. I also strengthen and extend this view by considering issues of poor judgement and weakness of will, and when and how we can be considered responsible for them.

ACKNOWLEDGEMENTS

I would like to thank all of the instructors I have had in the Philosophy Department over the past year, who have helped to create a very positive experience for me as a graduate student. These instructors – Paul Thagard, Steve Weinstein, John Turri, and Mathieu Doucet – have all helped me mature intellectually.

I would also like to express my gratitude to everyone else in the Philosophy Department. Every member of the department has contributed in some way, however small, to my great learning experience as a graduate student. I am grateful to both Paul Thagard and Patricia Marino for agreeing to be my readers, and for their excellent comments. Most of all, I would like to thank my thesis supervisor, Mathieu Doucet, who has been a constant source of information and guidance, and has had an immeasurable influence on the thought which has gone into the present project.

Tommy Blanchard, University of Waterloo

Table of Contents

Introduction....1
Chapter One....6
Introduction....6
A Review of Free Will and Responsibility....7
Types of Compatibilism....10
Moderate Reasons-Responsiveness....13
The Agent....15
True Self and Whole Self....18
Conclusion....23
Chapter Two....25
Introduction....25
Deep Self....26
Deep, Whole Self....29
Conflicts of desire....31
Connections and Hard Cases....33
Assigning Praise and Blame....36
Faulty Mechanisms and Responsibility....38
Conclusion....42
Chapter Three....44
Introduction....44
Clarifying Concepts....45
Self-Control....47
The Neuroscience of Self-Control....48
Ego Depletion....49
Self-Control and Responsibility....51
Judgement....54
Judgement and Blame....56
Judgement and Pathologies....61
Conclusion....64
Summary and Closing Thoughts....66
Bibliography....68

Introduction

Free-will is a question which many philosophers and lay-people alike have spent much time pondering over. This is for good reason. Free will is a deep issue which is relevant to our everyday experiences, our views of how the universe works, and our notions of responsibility. Views of free will have the potential to challenge our views on legal as well as moral matters.

Much of the debate over free-will is on the issue of moral responsibility. It is largely due to the repercussions views of free-will can have on our views of responsibility that the free-will issue holds such great importance. If the behaviour of agents is deterministic, this seems – to many people – to undermine the idea that they can be held responsible in a deep way. Libertarians[1] hold that, although in a deterministic universe we cannot hold people ultimately responsible for their actions, agents are not wholly deterministic and therefore they can be held ultimately responsible for at least some of their actions. Hard determinists[2] also accept that determinism and moral responsibility (as it is currently understood) are incompatible, and conclude that, since the universe is deterministic, ultimate responsibility does not exist.

Others reject the idea that determinism is incompatible with the freedoms required for any coherent sort of responsibility. Compatibilists[3] view determinism as compatible with, if not required for, moral responsibility. This captures the attractive features of both hard determinism and libertarianism. It accepts the libertarian view that responsibility is not an illusion or a mistake, and it also acknowledges, as hard determinists do, that we are part of the physical world, working under the deterministic laws of nature. However, this raises many questions. How can we differentiate the actions of a person from that of a falling rock? Both are simply following the laws of physics. How, then, can one be held responsible if it injures another human being, but the other cannot? Since both are obeying

1 e.g. Kane, 1998
2 e.g. Weatherford, 1991
3 e.g. Fischer & Ravizza, 1998

the laws of physics, neither the person nor the rock can behave differently. This realization brings forth a powerful intuition in many people – that if it is the case that people cannot act differently, it is not right to hold them responsible for their actions.

The intuition that agents must have alternate possibilities open to them in order to be held responsible, in its most naive form, has been powerfully refuted. Harry Frankfurt[4] has argued persuasively that there exist counterexamples which show that alternate possibilities are not required for responsibility. Although this has far from closed the debate over alternate possibilities and responsibility, it does raise the interesting possibility that our intuitions about responsibility are tracking something other than whether agents have alternate possibilities open to them. The compatibilist's job, then, is to find an account of responsibility which satisfies these intuitions in a coherent way.

There are many views of what is required for responsibility within the compatibilist camp. The current project is to review some of these views, using them to develop and defend my own view. This view posits that praise and blame are social attitudes[5], which have identifiable norms governing when they are applicable. I argue that in order for an agent to be held morally responsible for an action, the agent must meet two criteria. First, the agent must be able to properly understand and react to reasons for action. Second, the desire to act in a particular way must stem from the acting agent's self in the appropriate way.

Being able to recognize and react to reasons is called reasons-responsiveness. To be more explicit about what this means, an agent is said to be (at least weakly) reasons-responsive if there is some possible world where the facts relevant to a decision have changed, but the agent has not, and this results in a different choice by the agent. This idea seems to capture some of our intuitions about what sorts of agents have moral responsibility, and has been defended as the defining feature of morally

4 Frankfurt, 1969
5 As argued by Strawson in P.F. Strawson, 1962

responsible agents by Fischer and Ravizza.[6] It is certainly the case that agents without any sort of reasons-responsiveness do not seem to be proper candidates for moral responsibility. An agent who acts the same way regardless of the facts relevant to the decision is much more like a rock than like a morally responsible agent. This, then, gives us a way of differentiating morally responsible agents from inert matter – although they both are simply following the laws of physics, only responsible agents are reasons-responsive. However, this leads to an interesting question – there seem to be many cases in which a reasons-responsive agent acts, but is being compelled to act in a particular way despite what they really want.

There are many different addictions and pathologies (such as Obsessive Compulsive Disorder) which cause agents to perform actions which, in many cases, we do not want to hold them responsible for. Even if these agents are reasons-responsive, and acting based on the desires which they have, they may not be responsible for some of their actions. There are numerous views on why this is the case. Generally, though, they all claim that the relevant factor in these cases is that the agents' actions do not stem from, or are not endorsed by, their selves in a relevant way.

The view I endorse about what actions an agent is responsible for is a combination of the Whole Self view of Arpaly and Schroeder,[7] and the Deep Self view outlined by Sripada.[8] On this view, an agent is responsible if their actions stem from values or desires which are a part of their Deep Self, or (synonymously) are well-integrated. Parts of the Deep Self tend to be enduring, desires within it tend to be general (as opposed to directed towards specific things), they tend to be reflectively endorsed by the agent, these traits are often central to the agent's self-conception, and are not generally in extreme conflict with other deep traits. Although these criteria may not be rigorous, they may be the best that can be done – these are, after all, criteria which we (by hypothesis) track and make social judgements

[6] Fischer & Ravizza, 1998
[7] Arpaly & Schroeder, 1999
[8] Sripada, 2010

about. Thus it should not be surprising that they are a little fuzzy around the edges – social life is complex. There are many cases in which our intuitions about the responsibility of the agent are ambiguous and fuzzy – if the fuzziness of our intuitions matches up with the fuzziness in this account, then the lack of a crystal clear delineation is a virtue of the account.

This project thus seeks to scope out the territory of moral responsibility, and offer a critique of many of the current views on the issue. It will then put forward a particular view, drawing upon the insights of many of the other accounts on the issue. I will draw on empirical work in an attempt to provide some additional grounding to this view. It will also address some possible difficulties, and attempt to clarify some of the concepts at work in the background of this (and many other) views of responsibility.

The present project is broken into three chapters. The first chapter acts to further introduce the issues involved with moral responsibility, briefly mention some of the history behind the issue, and give a more in-depth analysis and critique of the various views on the issue. It explains the concept of reasons-responsiveness, how this helps to define morally responsible agents, and points out some of the limitations of this view – specifically, that it does not give an adequate way to differentiate motives for action that the agent is responsible for from those that the agent is not. I then look at several views which could fill this gap, offering a critique of many of these views, while defending Arpaly's Whole Self view. It closes with some residual questions which the Whole Self view leaves us with, most importantly, how to define what parts of agents we can view as a part of their Whole Selves.

The second chapter picks up the questions which the first chapter left us with. The Deep Self view is introduced and compared to the Whole Self view, in order to combine and strengthen both. Both of these views posit that some traits of our selves are in some way more representative of who we truly are, and it is only for actions stemming from these deep traits that we can be held responsible for. By comparing, critiquing, and combining these views, I am able to give additional clarity to the issue of

what constitutes a deep part of the self. The chapter then goes on to explicitly point out situations in which agents are not responsible for their actions – in particular, when their actions stem from faulty cognitive mechanisms instead of from their Deep Selves.

The final chapter discusses explicitly some of the concepts relevant to assessments of action which are often left implicitly assumed, and discuss some of the interesting issues that these concepts raise. Specifically, I discuss the scientific literature on self-control and judgement in an attempt to better understand these concepts, and raise the interesting questions of whether we can be held responsible for a lack of either. I argue that in some cases, we do hold people responsible for a lack of judgement or self-control, and that the combined Whole/Deep Self account can account for this.

Chapter One

Introduction

A large part of the free-will debate is over the question of moral responsibility. Is moral responsibility possible in a deterministic world? What are the necessary and sufficient conditions for an agent to be morally responsible for a particular action? Both of these questions have received wide attention in the philosophical literature. A number of different positions have sprung up out of this debate, giving a wide range of possible answers to these questions. In this chapter, I will briefly explain some of the positions, before going into more detail on one particular view – the reasons-responsive view.

Libertarians[9] hold that free will and ultimate responsibility are incompatible with determinism. In their view, one of the requirements for responsibility is the ability to do otherwise in a metaphysical sense. Since, under determinism, there is only one possible course of action an agent may take, libertarians see ultimate responsibility as requiring some amount of indeterminism. Compatibilists, however, disagree – they do not believe that a significant aspect of responsibility requires indeterminism. Compatibilists instead look for other requirements for responsibility.

There are numerous compatibilist views which claim different requirements must be met in order for an agent to be responsible. One such view is the reasons-responsive view. The reasons-responsive view holds that an agent is morally responsible if that agent is responsive to reasons.[10] That is, if there is some possible world where the facts have changed, but the agent has not, and this results in a different choice by the agent, we can say that this agent is (at least weakly) reasons-responsive. This is a powerful idea, and seems to capture some of our intuitions about what sorts of agents have moral responsibility. However, as outlined, the view is incomplete – what is required is a manner of

[9] e.g. Kane, 1998
[10] Fischer & Ravizza, 1998

saying when an agent is responsible for an action, as opposed to something else which happens to be controlling their body.

There are many cases of pathologies (such as Obsessive Compulsive Disorder) and addictions which cause actions which – though the agent may be acting in a way which would be responsive to certain reasons – we feel that the agent should not be held fully responsible for. One way to account for these sorts of cases is to give an outline of what is required for us to say that an action stems from the agent in a relevant sort of way. In an important sense, addictions and pathologies are often not integrated into a person's psychology, and are a sort of foreign object, forcing the agent to act in one particular way or another against his/her will. The Whole Self view will be presented as a way of separating well-integrated desires from those that are not well-integrated, and thus gives us a way of differentiating those who act on a desire and should be held responsible for it, and those who act on a desire and should not.

A Review of Free Will and Responsibility

Free will is possibly the philosophical question which has generated the most debate.[11] The importance of the debate is quite apparent: our notions of freedom are tied closely to our notions of responsibility. We tend to think it inappropriate to hold people responsible for things they have no control over. It is therefore of importance to ask: “Do we have control over our actions?” The worry is that, in a deterministic universe, the answer to this question may be a simple “No.” If all of our actions are determined by the physical make-up of the universe and the laws of physics, there does not seem to be much sense to the claim that we are free to do otherwise – there is only ever one course of action available to us.

Determinism is a central worry within the free-will debate. There are those (the libertarians) who argue that free-will and moral responsibility require some amount of indeterminism, and argue for

[11] Matson 1987, A New History of Philosophy. Vol. 1. New York: Harcourt, Brace, Jovanovich. quoted in Kane, 1998

a certain level of indeterminism.[12] Libertarians tend to focus on the role that indeterminism may play, and how this may allow for agents to have the ability to do otherwise. Hard-determinists agree with the libertarians that free-will and moral responsibility require indeterminism, but argue that, in fact, determinism holds and thus we do not have free-will and cannot legitimately hold people morally responsible.

A intuition which plays a large role in motivating this discussion is the principle of alternate possibilities. This principle states that in order to be morally responsible for an action, the agent in question must have been able to do otherwise. Since, in a deterministic universe, agents are not able to do otherwise, agents in such a universe cannot be morally responsible. However, this principle has been called into question. Harry Frankfurt gives the following counterexample:[13] imagine that Jones is deciding whether or not to pursue a particular course of action. Unknown to Jones, Black wants Jones to perform a specific action. Black is observing as Jones is deciding whether to perform the action or not, and, if he suspects Jones is going to do otherwise, will then act, manipulating Jones' brain in such a way to cause Jones to perform the action regardless. In other words, though he does not know it, Jones is unable to do otherwise. Yet, if Jones decides to perform the action without Black intervening, it is clear that Jones is morally responsible for this action. The fact that Black would have stepped in and changed things if Jones decision came out otherwise does not change Jones' responsibility in the matter.

There has been a long debate about what exactly these sorts of counterexamples mean for the principle of alternate possibilities.[14] Regardless, the argument does strongly suggest that what is of central importance to moral responsibility is the source of the action, and it is not clear whether this source must be indeterministic for moral responsibility to hold. Libertarians and hard-determinists, since they agree that determinism is incompatible with moral responsibility, share a position called (fittingly enough) incompatibilism. The opposing camp, compatibilism, holds that, in fact, determinism

[12] Kane, 1998
[13] Frankfurt, 1969
[14] For a review of this literature, see Levy & McKenna, 2009

and moral responsibility can co-exist. On this view, it does not matter whether we have the ability to do otherwise in a metaphysical sense, but the source of our actions must be of a particular sort for us to be morally responsible. That is, even in a completely determined universe, we could still have legitimate reason to hold agents morally responsible, as long as the action-causing mechanisms within those agents fit certain criteria. The compatibilists are in the business of trying to develop what those criteria are, and argue for how they might work to give rise to moral responsibility. They are generally of the opinion that all intelligible forms of freedom can be had, even if in some metaphysical sense all of our actions are predetermined.

The compatibilist project is to show how the resources available in a deterministic setting are sufficient to grant some freedoms which are sufficient for responsibility. Incompatibilists can recognize and appreciate some types of compatibilist freedoms, even if they disagree with compatibilists that these freedoms exhaust the freedoms we want and are sufficient to grant us responsibility. As Robert Kane says:

> [E]ven if we lived in a determined world, we could meaningfully distinguish persons who are free from such things as physical restraint, addiction or neurosis, coersion or political oppression, from persons not free from these things, and we could allow that these freedoms would be worth preferring to their opposites even in a determined world.[15]

Kane, a libertarian (and therefore an incompatibilist), is of the opinion that compatibilist accounts of freedom are important, and have been meaningful contributions to philosophy. His opinions differ from compatibilists in that he contends that the compatibilist freedoms do not give us all of the freedoms and responsibility we want, and that indeterminism can add something further.

For the purposes of the present project, I wish to remain neutral on the topic of indeterminist free-will. It seems it is perfectly possible to do work on a compatibilist project in such a way that does not preclude the existence or significance of indeterminism. I will not comment further on incompatibilist free-will here – it is not the topic I wish to pursue, and it suffices for me to state that the

[15] Kane, 1998, p 15

sorts of compatibilist notions of moral responsibility I will be discussing are consistent with determinism or indeterminism, and do not presuppose the presence or absence of indeterminist free-will.

Types of Compatibilism

David Hume is credited with articulating the most influential compatibilist position.[16] According to Hume,[17] an action is free if it is caused by the acting agent's willing. That is, if an agent, based on their desires, wills to perform an action, and in so doing performs the action, the agent acted freely. In this way, an involuntary twitch is different from a voluntary movement – the former was not willed and just happened automatically, while the latter was willed and caused by the agent's wanting to do so. According to Hume, what matters is not that we have, in a metaphysical sense, other possibilities open to us – it is that we are free to perform the actions we want to perform. Thus, when we act on a desire, we are responsible for that action. When we act because we were forced in some way, we are not responsible – our action does not stem from a desire to act in a particular way, but from an external force acting on us.

While Hume's view is extremely important, it is important to note that it seems to have counterexamples. Take, for example, unwilling addicts. They may have tried their best to quite the object of their addiction numerous times, and failed. Yet, they have a strong desire to use the substance, and in a sense willingly do so. Similarly, people suffering from Obsessive Compulsive Disorder have strong desires to perform their compulsions. Yet in these cases, it does not seem right to hold the agents responsible for these actions. In an important sense, it seems they are not free. Though they might, in a sense, be acting on what they want, it is clear that they are being forced. Desires can arise in and cause actions in a way which undermines responsibility.

One possibility which has been explored is the idea that not only must our actions line up with

[16] Russel, 2008
[17] Hume, 1748

what we want, but we must have a second-order desire for that want. That is, we have to desire to have the particular desire we are acting on. This line of thinking has been taken up by Harry Frankfurt, according to whom it is when we are able to want what we want to want that we are free.[18] In other words, when our higher order desires line up in the appropriate way with our lower order desires, we are free. You are morally responsible for an action if you did not have a higher-order desire to do otherwise. Unwilling addicts have desires for the substance to which they are addicted, but also do not want to have that desire. That is, unwilling addicts would, if they could, eliminate their want of the addicting substance (in contrast to willing addicts, who would not). Unwilling addicts thus have a first-order desire to take a substance, but a higher-order desire to abstain from it.

One could question why Frankfurt feels higher-order desires are so important for moral responsibility. If one has a first-order desire for a particular action, and a second-order desire to do otherwise, why hold the agent responsible for the second-order but not the first? The answer Frankfurt gives to this is that agents must identify with their first-order desires in order to be responsible for them. To identify with a first-order desire is to judge that any further questioning of that desire would lead to the same conclusion. In this sense, higher-order desires can be seen as double-checks on the initial answer, to see if the agent truly wants the target of the first-order desire. If these double-checks produce the same result, the agent is likely to decide that any further checking will also produce the same result, and thus the desire is one the agent truly wants.[19] This form of identification will be discussed again briefly below.

Peter Strawson has taken a different view on what is required for responsibility. He has argued that what grounds moral responsibility is our reactive attitudes.[20] The reactive attitudes are the attitudes and opinions we hold towards others in response to their actions. They are “natural human reactions to the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions.

[18] Frankfurt, 1971
[19] Frankfurt, 1987
[20] P.F. Strawson, 1962

[emphasis in original]"[21] As Watson points out, on this view, it is not that we hold people responsible because they are responsible, but that "the idea (*our* idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but of expressing our concerned and demands about our treatment of one another. [emphasis in original]"[22] In Strawson's view, the reactive attitudes are not things to be explained in terms of how we come to be responsible, but an inescapable, arational part of what it means to be human and to have social relationships.

On Strawson's account, we have reactive attitudes towards people when they are the proper beneficiaries of such attitudes. We do not hold certain people responsible (e.g. children or those with deep-rooted psychological abnormality) insofar as we see them as excluded from ordinary adult human relationships which spawn the reactive attitudes. As Galen Strawson points out, though, our intuitions about who is a deserving subject of the reactive attitudes can change if we (for example) view their actions as having come about by way of deterministic processes.[23] That is, the incompatibilist's worry seems to still live on in what sorts of explanations of actions can temper our reactive attitudes. It is not enough simply to state that those who we hold reactive attitudes towards are responsible. The Strawson account is, at best, incomplete, and we need a fleshed out theory of what our intuitions are tracking when we see someone as being either morally responsible or not morally responsible for a particular action. Many compatibilist theories, including the Frankfurt account outlined above, can be seen as attempts to achieve this fleshing out.

As the Frankfurt examples make clear, it is not simply being able to do otherwise which matters for moral responsibility. Yet, there seems to be some powerful intuition that something like the ability to do otherwise is important for being able to do otherwise. One possibility that has been explored is that it is the ability to do otherwise *if the reasons for acting have changed* that leads to moral responsibility. This position is called reasons-responsiveness. Reason responsiveness is perhaps the

[21] P.F. Strawson, 1962, p 53
[22] Watson, 1987, p 121
[23] G. Strawson, 1986

most common framework compatibilists work within. According to philosophers in the reason responsiveness camp, it is the ability for us to recognize and act for reasons which lays the foundation necessary for responsibility. This camp holds that as long as agents' actions stem from their own reasons-responsive mechanism, it does not matter whether these reasons-responsive mechanisms, the agents, or the world, are deterministic. What matters is that if the reasons were to change in a specific way, the actions of the agent in question would change. This has the advantage of requiring a specific sort of processing for moral responsibility, while still requiring agents to be able to "do otherwise", with the caveat that they be doing otherwise due to reasons being otherwise. With this in place, we can speak of the sorts of reasons for which an agent has acted, and whether these were morally praiseworthy or blameworthy. The view I wish to pursue lies in this camp, and thus this will be discussed in more detail below.

Moderate Reasons-Responsiveness

The view which will be developed here will be in the reasons-responsiveness camp. The remainder of this chapter will be fleshing out what seems to be, at least, a plausible account of some of the ingredients required to hold an agent morally responsible. This will provide the framework with which to work out the details of the mechanisms required for one to be a moral agent. I begin with the view of Fischer and Ravizza, who contend that what is necessary for moral responsibility is moderate reasons-responsiveness.[24] This amounts to three requirements. First, the agent in question must be strongly receptive to reasons. Secondly, the agent must be weakly reactive to reasons. Finally, and most simply, one must be able to act based on their choices.

Being receptive to reasons simply means being able to see and judge reasons for what they are. Being weakly receptive to reasons means that there is at least some imaginable set of circumstances which the agent could recognize as being reason to act in a particular way. That is, the agent is capable

[24] Fischer & Ravizza, 1998

of recognizing at least one possible scenario as being reason to take a particular action. This, Fischer and Ravizza argue, is not sufficient for moral responsibility.[25] The issue is that one can be receptive to reasons in an irrational or unpatterned way. If you find that tickets for a basketball game being $100 is sufficient reason to not buy a ticket, you should recognize tickets being $200 to be sufficient reason to not buy a ticket. Failure to do so impinges on your credibility as a moral agent – it seems that some base-line of objectivity in judging reasons is required for one to be morally responsible. Thus, Fischer and Ravizza suggest that what is required is strong reasons receptivity – your hierarchical ordering of reasons must be understandably patterned in an objective way. Thus, a third-party who asks you what actions you would take in particular circumstances would be able to understand the pattern by which you judge reasons.

The second requirement for being moderately reason-responsive is that the agent must be at least weakly reason reactive. While reasons-receptivity deals with the recognition of reasons, reasons-reactivity is the ability to choose a course of action based on reasons. Being weakly reasons reactive means that there is some set of reasons which you would find sufficient to do otherwise. Thus, without changing anything about the agent, there must be some set of circumstances which would cause the agent to choose to do otherwise. The reason that one only needs to be weakly reason responsive, according to Fischer and Ravizza, is that "if an agent's mechanism reacts to *some* incentive to (say) do other than he actually does, this shows that the mechanism *can* react to *any* incentive to do otherwise."[26] As moral responsibility already requires strong reasons-receptivity, the agent must understand and rationally judge reasons, and thus, according to Fischer and Ravizza, showing that they are capable of choosing one set of reasons shows that they could choose others.

The last criteria, that an agent must be able to act based on their choices, is the simplest. Quite simply, it just means that the outcome of the reasons-reactive mechanism in the agent must be connected in the appropriate way to action. If one is completely unable to act in accordance to their

[25] Fischer & Ravizza, 1998, pp 70-73
[26] Ibid., p 73

choice, they cannot be held morally responsible.

The Agent

A possible objection to the moderate reasons-responsiveness view is that it too readily assigns moral responsibility. In some cases, if one is perfectly able to understand reasons, but only capable of reacting to the most extreme, it seems callous to assign responsibility. Since only *weak* reasons-reactivity is required, on the Fischer and Ravizza view, all that is required of a morally responsible agent is that there be some conceivable set of circumstances in which the agent will choose to do otherwise. Yet for many cases, someone may be unable to choose to do otherwise in any situation likely to arise, but may be willing to act otherwise if a very extreme, unlikely set of circumstances is to arise. If an account holds agents responsible in these sorts of cases, this suggests that the account is wrong.

Take, for example, an extreme agoraphobic who can only muster up the courage to leave his/her house if he/she is firebombed, or something worse. That is, there is a conceivable set of circumstances which this agoraphobic would respond to, which means that he/she is weakly-reasons responsive. As long as the agoraphobic is willing to also leave his/her home if something worse than firebombing occurs (for example, if their home is fire bombed as well as filled with poisonous gas), then the agoraphobic has an objectively coherent hierarchy of reasons, meaning that he/she is strongly reasons-receptive. This makes this individual moderately reasons-responsive – he/she fit the criteria needed, by Fischer and Ravizza, to be held morally responsible. Yet, if this individual has made many honest attempts to get over the phobia and have been unable to, it seems wrong to place blame on him/her for not being able to leave their houses to maintain employment, or attend a wedding. Extreme situations often bring us to be capable of things that we would not otherwise be able to do. Sometimes people who are in extreme, life-threatening situations are capable of performing feats of seemingly super-human strength, lifting much more than they normally would be able to. Yet we do not expect these same individuals to be able to perform these acts in normal circumstances. In the same way, we should

not hold one responsible for failing to act a particular way just because in an extreme set of circumstances they would not fail.

This issue was raised by Mele in response to Fischer and Ravizza.[27] In replying to this worry, Fischer and Ravizza differentiate between moral responsibility and moral praise/blameworthiness.[28] In their view, to hold someone to be morally responsible is not to say that they are morally blameworthy or praiseworthy – it is just to say that they meet some standards of being responsible in some sense. Thus, though they would not want to claim that the agoraphobic is morally blameworthy for failing to leave the house, they would want to say that the agoraphobic is morally responsible for it, even if no moral judgement is made based on that action. Mele has pointed out that this is a strange use of terms. Though there is a difference between one who can do otherwise in at least the most extreme cases, and one who can not do otherwise no matter what (e.g. an agoraphobic who would rather burn to death than leave the house), it seems odd to use the term "morally responsible" to denote the former.[29] By using the term "moral responsibility" in this way, it seems the importance of a theory of moral responsibility has been diminished, and thus the importance of the Fischer and Ravizza model has been diminished.

A different tactic can be taken to deal with cases of individuals who are weakly morally reactive but do not seem to be morally responsible. One could claim that what these examples illustrate is that it is not enough to specify that an agent must be moderately reasons-responsive. We also need to worry about what part of the agent is doing the controlling. In cases of coercion, there is a powerful force outside of the agent which is forcing the agent to act in a particular way, and thus we do not hold the agent responsible. Yet there are also cases of internal forces which can completely override the agent's normal decision making procedure, and force the agent to act in a particular way. If the part of the agent which is doing the controlling in these cases is not central to the agent, then, much like the coercion cases, it seems the agent is being forced by something else to act in a particular way. In the

[27] Mele, 2000
[28] Fischer & Ravizza, 2000
[29] Mele, 2006

agoraphobia case, it seems the phobia is an affliction that the agents, in some sense, cannot help – they have made effort to rid themselves of it, and it seems like they are at the mercy of this strange, foreign phobia.

The Fischer and Ravizza view holds that the mechanism which is reasons-responsive must be the agent's in some sense. This is a possible way of dealing with the agoraphobic issue – if what accounts for the agents' actions is the phobia of open spaces, and that phobia is, in some sense, not part of the agents, then we can abstain from holding them morally responsible in this case.

Fischer and Ravizza hold that what makes a mechanism one's own is taking responsibility for it.[30] According to the Strawsonian account, which Fischer and Ravizza are working within, agents are responsible because they are held responsible by others. Similarly, Fischer and Ravizza hold that to be considered responsible for a mechanism, the agent must be willing to accept that mechanism as their own and be held accountable for the good or ill produced by that mechanism. At some point in the agents' history, they must take responsibility for a particular mechanism for that mechanism to be considered part of what they can be held morally responsible for. This does not mean a verbal commitment must be made – just that certain beliefs and attitudes be held towards that mechanism, accepting that you are responsible for it. Thus, perhaps it could be said that the agoraphobic is acting on a mechanism which they have not taken responsibility for, and can therefore not be held responsible for.

There are two problems with this approach. First, it does not seem that the reasons-responsive mechanism Fischer and Ravizza have in mind should be able to incorporate or exclude things which give rise to desires (such as our innate biological drives like hunger, learned preferences, or in this case, phobias). The mechanism they have in mind is simply one which is able to recognize and respond when desires form part of a reason to take a particular action. This would explain why, in their response to Mele, they do not invoke this line of reasoning. Second, it is not clear exactly what is involved in

[30] Fischer & Ravizza, 1998, chapter 8

taking responsibility for a mechanism, so it remains questionable how useful this concept is. However, there may be a way to maintain the taking responsibility account while admitting the utility of other accounts. Taking responsibility for a mechanism in this way is a historical account of responsibility – that is, it depends on events that have happened in the past which one has no access to in the present. Most of our judgements about moral responsibility, however, occur without seeming to know whether at some point in the past the agent we are judging has taken responsibility for the mechanism the action stemmed from. We are able to hold someone morally responsible without knowing their full history regarding their reasons-responsive mechanism – it would be an odd world indeed if we were unable to do so. Even if the account of Fischer and Ravizza is right, this shows that there are some facts about the world in the present by which we can approximate which mechanisms the agent has taken responsibility for, and it is worth looking for what these facts are. From within the Fischer and Ravizza framework, the other positions on this issue can be seen as tracking these facts about the world which give information about what mechanisms the agent has taken responsibility for. Fleshing out these other positions may then be more informative than the Fischer and Ravizza picture, even if Fischer and Ravizza are ultimately right that moral responsibility requires agents to take ownership over their reasons-responsive mechanism.

True Self and Whole Self

There are many other views on which parts of a person count as part of the morally responsible agent. We have already seen Harry Frankfurt's view, where a higher-order desire counts as part of the morally responsible agent if the agent decisively identifies with that desire. In this case, the agent identifies with a higher-order desire, privileging it above the others. Stump has proposed a revised Frankfurt account, in which it is the actions we are responsible for are those which stem from desires our intellect sides with.[31] This is a sort of identification of the agent with the intellect, again privileging

[31] Stump, 1993

this part of the agent as being the part relevant for moral responsibility. Others defend similar views, such as Dworkin[32] and Neely.[33] The common thread throughout these views is the privileging of one part of the agent over all others, making one part of the agent the seat of responsibility.

These views which privilege one part of the agent over others have been dubbed 'True Self' views.[34] They see one part of the agent as more centrally the self than the other parts, and privilege it above the others when assigning responsibility. It is certainly plausible that parts of us are more important and more central to our identity than others, and therefore more important for responsibility. Certainly this seems to be the case when considering agents with certain pathologies. Yet, to claim that one particular faculty or type of desire is *the* part which is central to an agent's identity and responsibility is to oversimplify. Human beings are messy creatures. To privilege one particular part as being the center of responsibility seems artificial – we are more than any single part of our selves.

In contrast to these views is the Whole Self view. The Whole Self view does not privilege any particular part of an agent. Instead, it posits that desires and beliefs which are not integrated into the person's psychology are to be considered as outside forces acting upon the agent. Those beliefs and desires that are integrated are part of the self. Regardless of whether the intellect or higher-order desires affirm an integrated desire or not, it is a part of the agent's self and therefore the agent is responsible for any actions stemming from it. While higher-order desires and the judgement of the intellect may tell us something about how integrated a desire is, they are not necessary for responsibility. A self-deceived businesswoman who thinks she is acting on the most honest of desires, but unconsciously is driven by greed, is responsible, since her actions are stemming from a part of who she is – a greedy human being.

There is an interesting type of counterexample to the True Self views which help to highlight the advantages of the Whole Self view. Arpaly brings up cases of inverse akrasia, pointing out that it is difficult for hierarchical theories to account for responsibility in these cases.[35] Akrasia is performing an

[32] Dworkin, 1970
[33] Neely, 1974
[34] Wolf, 1990
[35] Arpaly & Schroeder, 1999

action despite judging it to not be the best available option (all things considered). Inverse akrasia is performing an action which one judges (wrongly) to not be best. Thus inversely akratic actions form a proper subset of akratic actions – those in which our judgement is wrong, but our desires seem to be tracking the better option.

> An often cited example of inverse akrasia is Huckleberry Finn.
>
> > As Huckleberry becomes the friend of Jim, a runaway slave, his conventional southern moral convictions tell him clearly that he should proceed to return the slave to his lawful owner. He knows, so he believes, what the right thing to do is. To his embarrassment, however, Huckleberry finds himself psychologically incapable of doing what he believes to be the right thing. When an opportunity comes to turn Jim in, he feels too sick at heart, and displays what he takes to be weakness of will. He just cannot do it. Eventually he completely gives up the idea of turning Jim in, and consequently decides that he is a weak, bad boy, and that being moral is far too hard and thankless a task.[36]

It is debatable to what extent Huck is morally responsible in this case. On the one hand, he judged that what would be best is if he turned in his friend Jim. On the other hand, he did not act on this judgment due to his sympathy for Jim. What is clear, though, is that the True Self theories have a difficult time capturing this difficulty in assigning Huck (or others in similar circumstances) blame or praise. If Huck is responsible for the judgements of his intellect or for his high-level desires, it seems we should only be looking at his judgement that he should turn Jim in, and disregard that it was his sympathy which kept him from doing this. This seems to oversimplify the case.

The Whole Self View, put forward and defended by Arpaly and Schroeder,[37] avoids many of the issues the True Self views run into. This is a view which does not hold the self to be any particular part of the agent, but rather, holds that the degree to which a desire or value is psychologically integrated is what determines whether the agent is accountable for acting upon it. Desires, regardless of whether they are first-order or higher, can stem from deep within a person's psychology, and thus be just as much a part of them as their conscious choices. On this account, what matters is that desires be well-integrated into the agent's psychology. Thus, this view can better handle the ambiguity of the Huck

[36] Arpaly & Schroeder, 1999, p 162
[37] Ibid.

case. It is difficult to blame Huck for his conviction that turning his friend in is the right thing to do, because his qualms with taking this action may stem from deep within his character.

To be more specific, Arpaly and Schroeder hold that a belief or desire are well-integrated to the extent that "(1) they are deep; and (2) they do not oppose other deep beliefs or desires."[38] They go on to explain what is meant by this:

> A belief or desire is deep insofar as it is a powerful force in determining the actor's behavior, deeply held, deeprooted. Deep beliefs tend to resist revision (one's belief that one speaks a language is thus very deep; one's belief that there is still half a jug of milk left is much shallower) and deep desires tend to be satisfied with preference over shallower desires in contexts where a choice is forced (the desire to give emotional support to one's lover is deeper than one's desire to eat a healthy lunch each day, even though these desires might rarely or never in fact conflict). Beliefs and desires oppose each other when they can't all be true (beliefs) or satisfied (desires).

In the case of Huckleberry, Huck does not have a history of being motivated by racism, and his racist ideology (passed down by authority figures) is opposed by his sympathy, which seems to go much deeper. The depth of his racism and sympathy are up for debate, which leads to some room for ambiguity about this case.

To better illustrate the Whole Self view, take the cases of Lana and Greg.[39] Both Lana and Greg are kleptomaniacs – that is, both have urges to steal items of trivial worth, not for their worth or utility. These urges are not explainable by any need or desire other than the urge to take the items. Both Lana and Greg believe that stealing is morally wrong and find their urges embarrassing. Both want to stop stealing, and make efforts towards this goal. They differ in an important way, however. Lana has no positive feelings towards her stealing. Her urges make her life more difficult, and when she steals she feels guilty and often goes back to secretly return the item. Greg, on the other hand, has some positive feelings towards his stealing. Though he thinks stealing is wrong and is somewhat embarrassed by it, he often thinks about his stealing and smiles to himself about it. He is somewhat resentful towards respectable people, and he finds it pleasurable to think back to the people in the stores he steals from,

[38] Ibid., p 173
[39] Example taken from Arpaly & Schroeder, 1999

looking foolish as he operates behind their backs. He is attracted by risky and exciting endeavours outside of his stealing, and his friends (who do not know about his stealing) think that he ought to take life more seriously.

Of these two kleptomaniacs, it is hard not to treat them differently. In the case of Lana, whether we hold her somewhat responsible or not, it seems clear that in some sense she is not fully responsible for her actions. She seems to be a genuine victim of an affliction – she gains no pleasure from stealing, feels guilty about it, and even returns the items she steals. Her kleptomania seems in deep conflict with the rest of her values. Though we may be more sympathetic towards Greg than someone who steals out of greed or entirely out of malice, Greg seems much more at ease with his stealing than Lana. He gains some pleasure from it, though he has made attempts to quit he does not seem too genuinely upset by his stealing, and he seems to enjoy other similarly risky and irresponsible behaviours.

Arpaly and Schroeder, the originators of this example, attribute the difference in attitudes towards Lana and Greg to stem from the different degrees to which their kleptomania are psychologically integrated. Arpaly and Schroeder's concept of psychological integration seems like a plausible and useful concept with which to deal with these cases. However, there are many areas in the account which are troubling or require elucidation.

First, the explanation of when beliefs and desires oppose each other seems too weak. According to Arpaly and Schroeder, beliefs and desires oppose each other only when they are logically mutually exclusive. Intuitively, it seems that beliefs and desires should be able to oppose each other to various degrees – a belief may make others less likely to be true, or satisfying one desire may make others more difficult to satisfy. For example, a belief that a friend owns a house may make it seem less likely that this friend also rents an apartment, though it does not logically rule out the possibility; satisfying one's desire for alcohol may make satisfying one's desire to get work done more difficult, but not impossible. Unwilling addicts may technically be able to satisfy any of their desires in addition to their desire for drugs, it will just be made more difficult. This issue could fairly easily be solved by having

desires and beliefs able to oppose each other to various degrees – the strongest degree when they are mutually exclusive, and moving down the spectrum as it becomes progressively easier to satisfy both. This introduces another issue which needs to be resolved: we all have desires which oppose each other to some extent, which seem perfectly a part of our "Whole Self". We need a method of differentiating those desires which are just at odds with some of our other deep desires, and those that truly are so opposed by our deep desires that they are not truly a part of us.

A more serious problem with the Whole Self account is that, though we intuitively have some idea of what sorts of traits are integrated and to what degree, Arpaly and Schroeder do not give us a particularly strong definition of what psychological integration amounts to. We are told that a desire is well-integrated to the extent that they are deep and that they do not oppose other deep desires. While this is helpful and suggestive, the definition they give of "deep" is circular. A deep desire is "deeply held" and "deeprooted". While we are told two properties that deep desires have – that they tend to resist revision and they tend to be satisfied with preference over shallower desires – this, again, is merely suggestive, and not particularly explanatory. A much more satisfying definition of psychological integration may go a long way to strengthening the Whole Self account.

We are left with two connected problems with the Whole Self account. The first is that we need some method of differentiating a person with a normal amount of integration and opposing desires from someone with a desire which is truly foreign to the self. The second is that we do not have a satisfying definition of psychological integration. The next chapter will deal with these issues.

Conclusion

I have outlined some of the requirements for an agent to be morally responsible. Following Fischer and Ravizza, I suggest that an important part of being morally responsible is being reasons-responsive. I also suggest that the desires one is acting on must be an integrated part of the agent – that is, they must be well-integrated in the agent's psychology. Taken together, this gives a powerful view of

what is required for moral responsibility, leaving the question of indeterminism to the side. However, the Whole Self view has left us with some residual issues, which will be dealt with in the next chapter.

Two interesting, related issues have yet to be addressed.. The first is the role of self-control. Some cases of inverse-akrasia were briefly mentioned, but there are many cases of agents who fail to keep themselves from acting on a strong (but not overwhelming) desire which is not well-integrated. On the Whole Self account, acting on one desire over another shows, all else being equal, that one is deeper than the other. What can we say about self-control, and those who are particularly bad (or good) at it? It is difficult, with the view sketched thus far, to distinguish between someone who has a nearly irresistible, poorly integrated desire for sweets, and one who likes sweets and simply is terrible at controlling themselves. In many cases, it seems like we want to hold people responsible, even if they are acting on something that is not well integrated, because we think they should have been able to control themselves.

The second issue is the role of judgement. Sometimes, we judge poorly, which causes us to act in a way not fully stemming from a well-integrated part of ourselves. Can we be held responsible for poor judgement, and if so, under what circumstances? Both judgement and self-control will be discussed in chapter 3.

Chapter Two

Introduction

In the last chapter, I introduced the idea of the Whole Self, defended by Arpaly and Shroeder.[40] This view holds that an agent is to be held morally responsible for actions which stem in the appropriate way from some well-integrated aspect of their psychology. Thus, actions stemming from the addictions of unwilling addicts, or the compulsions of patients with Obsessive Compulsive Disorder, are not the sort of thing an agent can be held accountable for, since these actions stem from a poorly integrated desire within the agent. The Whole Self view left us with two related issues: The first is that we need some method of differentiating a person with a normal amount of integration and opposing desires from someone with a desire which is truly foreign to the self. The second, which is required to answer the first, is that we do not have a satisfying definition of psychological integration.

In this chapter, I introduce the Deep Self view, in order to bring some clarity to these issues. The Deep Self is similar to Arpaly and Shroeder's Whole Self – it is made up of the parts of the agent which really are central to that agent. By comparing and connecting the two views, I attempt to shed some light on the issues left by the Whole Self alone. The concept of the Whole Self will be brought into greater focus, but, I will argue, the concept is necessarily fuzzy. It is a folk-psychological concept, which people use in their attributions of praise and blame. Our attributions of praise and blame themselves are imprecise – there are plenty of cases which fall into some sort of grey area. We should therefore be wary of any attempt to offer a rigorous, precise definition of what these attributions are tracking.

After the discussion of when an agent *is* morally responsible for an action, I move into a discussion of when an agent *is not* morally responsible. I claim that one way for an agent to lack responsibility is if their actions stem from a broken cognitive mechanism. In many cases of

[40] Arpaly & Shroeder, 1999

pathologies, there are physiological problems – certain neurological mechanisms are malfunctioning. In these cases, it is the faulty mechanism which has caused the action, and thus we can rule out that the action was caused by some deep aspect of the agent's psychology, and thus the agent is not to be held responsible. These faulty cognitive mechanisms cause a disconnect between agents' Deep Selves and their actions, and serve as an external factor causing the agents' actions.

Deep Self

Chandra Sripada has put forward the Deep Self model,[41] which is in many ways similar to Arpaly and Schroeder's Whole Self view. The Deep Self view, like the Whole Self view, posits that attribution of moral responsibility is connected with the deep aspects of an agent's psychology. The Deep Self is made up of the agent's values, principles, and life goals. The parts of the Deep Self have many of the same characteristics as Arpaly and Schroeder's conception of well-integrated desires. For example, acting in accordance with them generally will take priority over satisfying less deep goals, and these personality traits are stable and unlikely to change. It seems plausible that the sorts of traits Sripada would be considered well-integrated traits on the Whole Self account. There are some characteristics of Deep Self traits which Sripada adds which the Whole Self view does not attribute to well-integrated desires. For example, Deep Self characteristics tend to be central to the agent's self-identity, and the agent will tend to reflectively endorse these traits – though this latter characteristic is one the Whole Self view fairly explicitly rejects as central. Attitudes which are part of the Deep Self also are more general than traits which are less deep.

Sripada gives some evidence for his claim that the traits he identifies as part of the Deep Self are central to ascriptions of praise and blame. Take the following scenario:

> The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’
>
> The chairman of the board answered, ‘I don’t care at all about harming the

[41] Sripada, 2010

> environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

Contrast this with the similar case:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was helped.[42]

It seems intuitive that in the first condition, we blame the chairman for opting for a program that will harm the environment. In the second, we do not credit the chairman for helping the environment. There is some evidence to suggest these intuitions are held by a majority of people.[43] Yet these two cases are very closely analogous – in both, the chairman is making a decision between pursuing a project or not, and the project is said to increase profits. The chairman makes it clear in both that he does not care about the environment, and so the environmental effects are just a side-effect of his attempts to increase profits. Why, then, is there an asymmetry in how we assign credit for the resulting effects on the environment?

Joshua Knobe, the originator of these vignettes, suggests the hypothesis that "People's judgements [about intentionality] depend in a crucial way on what [the action] happens to be. In particular, it makes a great deal of difference whether they think that [the action] is something good or something bad."[44] Knobe contends that the reason we assign intentionality to the chairman who harms the environment, but do not do so the chairman who helps it, because of the moral valence of the side-effect. The asymmetry between our attributions of intentionality in the two cases stem from an asymmetry in how we view good and bad actions. This explanation, however, has problems with the

[42] Both scenarios are from Sripada, 2010, p. 161

[43] Knobe (2003) reports that, of 78 people asked in Manhatten public park, 82% of people said that in the first condition, harm to the environment was intentional, while 77% said that the help to the environment was not intentional.

[44] Knobe, 2003, p. 190

follow examples:

> **Aunt Killer**
> Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger. But John isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet hits her directly in the heart. She dies instantly.
>
> **Selfless Soldier**
> Klaus is a soldier in the German army during World War II. His regiment has been sent on a mission that he believes to be deeply immoral. He knows that many innocent people will die unless he can somehow stop the mission before it is completed. One day, it occurs to him that the best way to sabotage the mission would be to shoot a bullet into his own regiment's communication device.
>
> He knows that, if he gets caught shooting the device, he may be imprisoned, tortured or even killed. He could try to pretend that he was simply making a mistake – that he just got confused and thought the device belonged to the enemy – but he is almost certain that no one will believe him. With that thought in mind, he raises his rifle, gets the device in his sights, and presses the trigger. But Klaus isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…Nonetheless, the bullet lands directly in the communications device. The mission is foiled, and many innocent lives are saved.[45]

In both these cases, most subjects (over 90%) presented with these scenarios say that the result was intentional. This calls into question the notion that it is the moral valence of the consequence which affects our attributions of intentionality, as Knobe posits.

Sripada explains this asymmetry by applying the Deep Self model he has developed: in cases where the effect of an action is in concordance with the person's Deep Self, we see the act as intentional, and then assign praise or blame. In both of the chairman examples, it is made clear that the chairman values profits over the environment. In only the first scenario was the effect concordant with these values – the environment suffered because of the chairman's values. We do not give the chairman credit for helping the environment in the second scenario, since the beneficial effect on the environment does not derive from the chairman's valuing of the environment. Although we may see the chairman as of a questionable moral character, there is no effect in the second scenario that we can blame him for. In the Aunt Killer and Selfless Soldier scenarios, the results of the actions are in line

[45] Sripada, 2010, p. 170-171

with the actor's Deep Self, and therefore the actor is judged as having intentionally brought about the result. Thus, it is when an agent is acting in a way that reveals their Deep Self that they are perceived as responsible for those actions.

Deep, Whole Self

Sripada does not give a precise definition of the Deep Self, or a method of discovering whether a particular value is part of the Deep Self or not. He is satisfied with an imprecise definition, precisely because he sees the Deep Self as a folk-psychological concept people use when assigning moral praise or blame. This makes sense. If our intuitions about responsibility are tracking something like the Deep Self, it could very well be a fairly fuzzy concept. This does not mean that the Deep Self is not a useful concept for differentiating different cases of potential moral responsibility – it just means that there may inevitably be some amount of imprecision in the concept. There are many cases where it is not clear that we should hold someone responsible, and we should be wary of a view which divides cases up with too much precision – a view of responsibility should make sense of why we are puzzled by marginal cases, but see central cases as clear-cut, and it seems the Deep Self may fit this bill. We may need to content ourselves with suggestive, but not rigorous, definitions of the Deep Self to explain our intuitions about moral responsibility. This may be why the definition of well-integrated, from the Whole Self view, is similarly fuzzy. Both of these accounts are able to give suggestions about the sorts of traits which we see as a core part of the agent, but neither can give a full definition.

One of the questions the Whole Self account left us with in the last chapter was what is meant by “well-integrated”. According to the Whole Self view, we are responsible for actions when the movitvation is well integrated. Similarly, the Deep Self view posits that we assign responsibility to agents for their actions if their actions stem from part of their Deep Self. The descriptions given of well-integrated desires and the traits associated with the Deep Self have much in common, and it seems these are just two different angles taken of the same concept. Henceforth, I will use the terms “well-

integrated" and "Deep Self" such that "well-integrated" is an adjective for parts of the Deep Self, and every part of the Deep Self is well-integrated.

I have suggested that though neither Arpaly or Sripada give a rigorous definition of the Deep Self, a rigorous definition may not be possible, and the suggestive nature of the two accounts may be the best that can be done. The best answer to the question of what "well-integrated" means may just be a suggestive list of characteristics. Some of the characteristics of well-integrated or Deep Self traits are:

- They are enduring
- They often motivate the agent to action, forming recurring patterns of behaviour
- They are general, not specific
- The agent tends to reflectively endorse them
- They are central to the agent's self-conception
- They are not in deep conflict with other traits

Again, these are suggestive, characteristic properties. None of them are necessary or sufficient, and this list is probably not exhaustive.

The first characteristic, that Deep Self/well-integrated traits tend to be enduring, is found in both the Deep Self and Whole Self account. This seems fairly intuitive. Though people do change, changes tend to occur over long periods of time. The deep aspects of a person's personality and values do not generally change over night. When a radical change seems to occur rapidly, we generally posit some special factor which has caused this – a mid-life crisis, or brainwashing from an ideologue.

The two views also agree that these traits generally have significant motivational force, and (since they are enduring) tend to form recurring patterns of behaviour. If an agent claimed to care very much about the environment, yet never seemed to act in ways to prevent environmental harm, we would question whether this really is a particularly deep value. An important characteristic which the Whole Self view does not mention is that deep values and desires tend to be general. Placing a high value on a particular, such as purchasing energy-efficient light-bulbs, is less deep than the more general

value of reducing one's energy consumption, which in turn is less deep than desiring to do less harm to the environment. The shallower, more specific desires are working in service of the deeper, more general. If it turned out that buying energy-efficient light-bulbs actually did more harm to the environment than good (for example, by producing much more pollutants in the manufacturing process than competitors), the more specific desire would evaporate. The deeper value, however, would be much more difficult to change, and seems much more central to what sort of person the agent is.

The next item, that the agent tends to reflectively endorse deep attitudes, is put forward by Sripada. Arpaly and Schroeder argue[46] that consciously endorsing one set of attitudes is not necessary or sufficient for deep attitudes – and this seems very likely to be the case, as cases like Huck Finn's show. However, it also seems that, in general, agents usually tend to endorse their core values when reflecting on them. Though there may be many counterexamples, this does not cast doubt on the general tendency for people to consciously agree with the values which make up who they are. This brings us to the second last item – deep traits tend to be central to the agent's self-conception. If it is true that, in general, deep attitudes are reflectively endorsed by the agent, it seems to follow that these values are also going to play a role in the agent's self-conception. Though there are many instances where we are deluded and have a distorted conception of our selves, values which we see as central to who we are are more likely to be deeply ingrained than ones we do not see as important to us.

Conflicts of desire

The final point is found exclusively in the Whole Self view. The claim is that a desire is well-integrated insofar as it is deep and not in conflict with other deep desires. Two desires are in conflict if it is not possible to satisfy them both or, as I argued in the last chapter, if satisfying one makes it more difficult to satisfy the other. However, this criterion seems controversial. It is obvious that in many cases – such as unwilling addicts and OCD patients tired of their compulsions – there is a conflict between the agents' deep desires and their addiction or pathology. However, it is also true that many

[46] Arpaly & Shroeder, 1999

desires, which seem to be very much a part of us, are in conflict. Most cases of weakness of will seem to demonstrate this. The dieter's desire to lose weight is constantly pitted against tempting sweets; yet, we do not want to claim that the dieter's desire for sweets is not well-integrated, since – according to the Whole Self view – this would mean that the dieter is not responsible for actions stemming from such a desire.

It seems that in these cases, there is sufficient integration to hold people responsible, despite conflicts. Yet if both the desire to lose weight and the desire for sweets are deep and well-integrated, we have an example of two deep desires which are in pretty serious conflict for which the agent is responsible. There are numerous other examples – the desire to do work is often in conflict with the desire to play; the desire to be generous is in conflict with the desire to accumulate wealth; the desire to be well-respected can come into conflict with the desire to be moral. Patricia Marino makes the case[47] that holding even essentially conflicting desires – wanting A and not-A – is not irrational, and there are many examples of such desire pairs which seem perfectly fine. For example, two friends with a friendly rivalry competing in a race may want, at the same time, their friend/rival to both win and not win. Though these desires are in essential conflict, they are understandable, and it does not seem that the conflict between them makes either one seem more external to the agent. These seem like values and desires which we want to say are a part of the Deep Self, and which, in general, we would want to hold an agent responsible for acting upon, despite the recognition of the conflicts.

As has been mentioned, none of the points are meant to be necessary or sufficient conditions. However, the issue with this last item seem concerning. If many of our deep values and desires are in conflict, it seems we are put in a position with two options: Conclude that conflict between desires and values is irrelevant (or at least, a very weak factor), or try to differentiate the sorts of conflicts which are acceptable within the Deep Self, and those which indicate a desire outside of the agent's Deep Self. The former option is disagreeable, since it seems that a very strong factor which curbs our blame of

[47] Marino, 2009

unwilling addicts (or patients with OCD or other, similar pathologies) is that they are unwilling – they have strong inclinations and values against continuing substance use.

There do, though, seem to be differences between the normal conflicts within each of us and the conflicts which occur within an unwilling addict. For one thing, there is a difference of extreme. While my interest in being healthy and avoiding empty calories is often in conflict with my sweet-tooth, this is hardly comparable to addicts who are unable to pursue many things of deep value due to their addictions dictating large portions of their lives. A person with a serious addiction may find it difficult to: maintain a job, maintain relationships and social status, pursue other desires which require financial resources, stay healthy, etc. Someone with a sweet-tooth, on the other hand, is only affected in a very limited number of ways (namely, that they will have trouble maintaining a balanced diet – an issue most of us face), and is unlikely to face temptations as difficult to resist as the drug addict's. Furthermore, many unwilling addicts seek ways to lessen or eliminate their addictions, by abstaining from drugs, attending addictions counseling, and/or going to drug treatment clinics. That is, the agent's deep values can sometimes lead to very costly actions for the sole purpose of attempting to eliminate the conflicting desire. It seems, then, that an important factor in differentiating normal value conflicts and the sorts of conflicts addicts face is to appeal to the strength of this conflict, and the strength of the deep values being opposed. When the conflict is extremely strong, especially when it leads to costly actions to eliminate one of the desires, it seems reasonable to say this becomes a factor (though again, not a sufficient or necessary one) against that desire being well-integrated.

Connections and Hard Cases

We can now test some of the above-listed characteristics of deep-self traits by pitting them up against some hard cases. In so doing, we will further refine the characteristics by drawing connections and relations between them. The issue with the characteristics as they currently stand is that many cases in which we would not want to assign responsibility seem to fit many of these characteristics. The

worry is that these characteristics may be an intuitive folk-theory, but does not actually demarcate cases of actions which stem from the deep-self, but rather just give some post-hoc rationalization for it.

Imagine a cocaine addict, Andy. Andy became addicted to cocaine as a teenager, when he was peer-pressured into using it on numerous occasions. Now, 20 years later, Andy is still addicted. His addicting has obviously been enduring, having lasted many years. It has also caused recurring behaviour – Andy needs to seek out and use cocaine on a fairly regular basis. The desire for cocaine is not particularly general – it is for a particular sort of substance – and Andy knows, on reflection, that this addiction is detrimental to his life. But the addiction has been a part of Andy's life for so long, it has become a part of his self-conception.

Andy's addiction definitely has 3 out of the 6 characteristics listed above – it is enduring, causes recurring patterns of behaviour, and is part of his self-conception. Yet he is also an unwilling addict, it seems plausible to say he is not responsible for his addiction – he began using cocaine as an immature teenager, and it can be said that in many ways he is a different person now. He would not have begun using cocaine if he had been a mature adult. Assuming that Andy has attempted to quit using the resources available to him and that he honestly still struggles with this addiction, it seems at least plausible that we should not say Andy is responsible when he succumbs to his addiction.

Are three characteristics enough to make a trait a part of the deep-self? Compare Andy to a miser, Mary. Mary is extremely greedy when it comes to money, but not with anything else – hence, her desire for wealth is no more general than Andy's want for cocaine. She simply likes the feeling of having wealth – to see large numbers in her bank account statement gives her pleasure, even though she does not wish to use this wealth for anything in particular. Her miserliness is not part of her self-conception – she rationalizes her actions towards money in such a way that it never occurs to her that she has any particular affinity for money. Her greed also rarely conflicts with her other deep desires – she does not place a high value on her living conditions or generosity. Although Andy and Mary both only have three of the six characteristics, it seems Mary is more responsible for her actions than Andy

is of his.

It seems, on the face of it, Andy's recognition of his addiction as part of himself does not really make us more likely to make him responsible – take this characteristic out of his story, and nothing seems changed. It seems that this is because he holds it as part of his self-conception while at the same time does not reflectively endorse it. He grudgingly accepts his addiction as part of who he is, but he wishes that it was not. In the terms of Frankfurt,[48] Andy has a first-order desire for cocaine, but his higher-order desires are in conflict with this desire – he does not want to want cocaine. Andy, on reflection, knows that his addiction is a bad thing. Not only does Andy realize his addiction is negative, but he is also serious enough in this judgement to have made attempts to eliminate it. His reflective judgement seems to take on a privileged position among the criteria – if his reflective judgement approves of a desire, considering it a part of his self-conception may help to solidify this trait as a part of his Deep Self, but without the reflective endorsement, or even reflective opposition, Andy's self-conception on this matter does not seem to change our evaluation of responsibility.

As noted in the previous section, conflicting desires alone does not mean that a desire is not well-integrated. However, certain sorts of conflicts do – when the conflict is of such a kind that it is strong and one is motivated to eliminate a desire, this speaks against the integration of that desire. This surely requires some amount of reflective opposition to the desire, as well as the opposing desires being motivating. Again, it seems reflective endorsement or opposition is important for deep-self traits.

Reflective endorsement seems to play a fairly central role. However, it is not as simple as claiming that reflective endorsement is the only characteristic required – as has been argued in the previous chapter, this over-simplifies. Frankfurt's heavy emphasis on reflective endorsement[49] stems from the fact that reflective endorsement is so important, but even in cases where the agent's reflective judgement is obviously against a desire, it may still not be clear whether the agent acting on it is responsible (see the case of Huckleberry Finn in the previous chapter).

[48] Frankfurt, 1987
[49] For example, in Frankfurt, 1987

Assigning Praise and Blame

In the last section, I combined the insights of the Whole Self and Deep Self views, and argued that what they are getting at may be complex with many interacting characteristics. I then listed a number of these characteristics. This, I suggest, may be the best we can do to answer the question posed in the last chapter about what it means for a desire to be well-integrated. I also attempted to suggest one way in which unwanted addictions differ from the regular conflicting values we all hold – cases of addiction seem on the extreme end of conflicting values. In this section, I want to look specifically at how the view I have been outlining might be used to assign praise and blame, and use this to illustrate the differences between cases where we do not hold agents responsible (such as in certain cases of addictions and pathologies) and cases in which we do.

We all, as human beings, make judgements about who is responsible for what. Any layman, completely uneducated in moral philosophy, is willing and able to judge whether a person is responsible for any particular action. Much of the time, these judgements are just based on simple facts, adduced or observed, about the case at hand – did the person know the consequences of the action? Was the action accidental, or was it intended? Could an alternative action have been taken? Etc. We are also able to pass judgements on the harder cases which have continually resurfaced here – cases of addictions and pathologies. Most people would agree that people suffering from severe Obsessive Compulsive Disorder, who spend many hours every day scrubbing their hands, are not responsible for failing to take advantage of their time in a more useful way.

What, then, is it that these everyday judgements are tracking? What is the common thread which binds all of these different cases together? In the previous chapter, we discussed the reasons-responsive view of Fischer and Ravizza, which holds that a person is responsible as long as their actions stem from their own reasons-responsive mechanism.[50] That is, if an agent is able to judge reasons in an appropriate fashion, and react to reasons, and these reactions come from a mechanism

[50] Fischer & Ravizza, 1998

which can appropriately be called the agent's own, then the agent is responsible. The trouble comes in defining what makes such a mechanism one's own, and dealing with the difficult cases where it is not entirely clear whether a particular reaction is coming from the agent or from something better conceived as outside of the agent. In the latter part of the first chapter, and in this chapter thus far, it has been suggested that agents are responsible for actions which stem from some part of their deep-self. This, however, gets this somewhat backwards. As was briefly pointed out in chapter 1, if we take something like the Strawsonian account,[51] then, as Watson has pointed out,[52] it is not that we hold people responsible because they are responsible, but that it is part of a human practice of holding each other accountable. That is, judgements about responsibility should be taken to be much more central to the notion of responsibility. Anything which seems to track our intuitions about responsibility tracks responsibility by virtue of lining up with our judgements, which hold the true nature of responsibility. Therefore, it is worth making clear the (perhaps obvious) point about how exactly we make judgements about responsibility.

The point is this: we hold someone responsible when we think that their actions reveal something morally important about their deep-self. If it seems that the action stems from the agent's ill will in some way – that is comes from the agent's selfish desires, or desire to inflict pain in others – then we consider them blameworthy. We also seem sometimes to hold people responsible when the outcome is caused by a failing, such as poor judgement or thoughtlessness. One possibility is that these cases actually reveal something deep about a person's values – the lack of effort put into planning, or lack of salience of a particular issue, may reveal something deep about an agent. Perhaps we also blame people for a lack of good judgement, because their lack of good judgement reveals a failure to cultivate good judgement, which may stem from a lack of a deep value for judging correctly. However, it seems like it may be special pleading to claim that this is the explanation for all such cases, and so perhaps we also hold people responsible for certain failings (particularly in judgement). Setting these possible cases

[51] That is, the account put forward in P.F. Strawson, 1962
[52] Watson, 1987

aside, the general principle is this: if an action reveals, or allows observers to predict, the deep-self of the actor, then the actor is held responsible (and thus is responsible). Note that the proper interpretation of “reveals” or “allows”, in the preceding statement, is required to disallow responsibility for actions which, though not stemming from a deep-trait, still lead observers to the correct conclusion that the actor has some particular deep-trait. So, for example, an evil genius bent on causing wide-spread misery may be wrongly held responsible for contributing to global warming through the releasing of some pollutant if the releasing of that pollutant was done without knowledge of the environmental effects. Although the releasing of this pollutant may lead observers to (rightly) assume that the evil genius is evil, upon learning that he had no knowledge of the effects of this pollutant, responsibility for this action wanes. There must be the proper connection between the deep trait and the actions of the agent – the predictive power gained by the observers must not be accidental.

Faulty Mechanisms and Responsibility

I have attempted to flesh out and connect the Whole Self and deep-self views in order to shed some light on what sorts of desires we are responsible for acting on. There is another way of approaching this issue, and that is to look for what sorts of desires we are not responsible for acting upon. This will serve to provide additional means to demarcate the desires we are responsible for acting upon, and those we are not.

I wish to here focus on brain mechanisms. In many of the cases which have been discussed thus far – addictions of various sorts and Obsessive Compulsive Disorder – there is something wrong in the agent's brain. That is, some mechanism is not functioning as it should. This requires some amount of caution, however. We do not wish to end up saying that “My brain made me do it” is an adequate moral (or legal) defense against responsibility. As Arpaly has said, “If the mental can be reduced to the physical (or materialism is otherwise true), then *all* mental states are, at bottom, physical states

[emphasis in original].”[53] So explaining that a particular action was caused by a physical circumstance in no way differentiates it from any other action. There is almost certainly something about a murderers brain which “made them” commit the murder, but this is not necessarily a defense against responsibility – if the only real difference between the murderer's brain and any pacifist's corresponds to a difference in personality, then this is not the sort of brain difference the murderer can appeal to in order to dodge responsibility.

The sort of issue that can mitigate or eliminate responsibility is one in which there is a legitimate problem in the functioning of some brain mechanism. That is, some brain mechanism is simply not working within the range of typical functioning. Take, for example, Obsessive Compulsive Disorder.[54] People suffering from this disorder are plagued by obsessions, thoughts which repeat over and over, despite being unwanted. More important for our purposes here, sufferers also carry out compulsions, behaviours which the subject generally knows are nonsensical, but feel compelled to do them regardless. Although the behaviours are generally fairly common behaviours that many of us may habitually do more often than strictly necessary, as personal quirks, in OCD patients, it is taken to the extreme, becoming pathological. People with OCD may, for example, have a compulsion towards hand washing. Though some people without OCD no doubt wash their hands more often than necessary, someone with OCD can spend hours in a day washing their hands, without feeling any behavioural closure to the matter. They continue to feel the need to wash, just as people with OCD who are “checkers” will continually check to see if, for example, their door is locked, and will be continually hounded with thoughts of what could happen if the door is unlocked, compelling them to check yet again.

OCD and its symptoms have been linked to three areas in the brain: the orbitofrontal cortex, the anterior cingulate cortex, and the basal ganglia. Subjects with brain lesions in these areas (both humans and animals) have been studied, and the results suggest that these three areas are implicated both in

[53] Arpaly, 2005
[54] Graybiel & Rauch, 2000

OCD patients and with OCD-like behaviour. Together, these three areas form what is sometimes called the "OCD circuit".[55]

The orbitofrontal cortex is associated with forming behavioural plans based on estimates of positive or negative consequences of particular plans.[56] When lesions are present in this area, it can lead to a mismatch between the expected outcome of a particular action and the behaviour of the subject – simply put, the mechanism which weights the value of different actions seems to be affected by these sorts of lesions. Thus, such a lesion may cause a feeling towards a specific action that is far out of proportion with the actual value of that action, which may lead to repetitive action. The anterior cingulate cortex is associated with motivation and affective behaviour. It is closely connected with the motor cortex, and plays a role in action selection. Together, these regions of the brain play a very important role in action selection: "As part of a cortical network, then, the orbitofrontal cortex and anterior cingulate cortex could together exert a powerful influence both on the perceived emotional value of stimuli and on the selection of behavioral responses based on these experience-based expectancies and perceived outcomes."[57]

The basal ganglia plays a role in habit formation. One hypothesis contends that it may not only play a role in motor habit formation, but cognitive habits as well. This hypothesis puts down OCD to a dysfunction of the loop between the basal ganglia and neocortex.[58] Instead of helping to form normal patterns, in a dysfunctional cortico-basal ganglia loop, pathologically recurring thought patterns (obsessions) are caused, along with repetitive actions (compulsions).

The neurobiology of OCD is far from being completely understood, but from what is known it is clear that there are certain neurological mechanisms which are not functioning the way they should – some form of lesion or neurochemical imbalance causes these mechanisms to malfunction. Patients with OCD often recognize their behaviours as irrational, but feel forced by something beyond their

[55] Graybiel & Rauch, 2000, p. 344
[56] Damasio et al, 1990
[57] Graybiel & Rauch, 2000, p. 344
[58] Ibid., p. 345

control to perform them. The compulsive behaviours individuals with OCD engage in are a product of bad brain mechanisms, they do not derive from higher-level concepts, such as values or personality traits. Though these are different levels of description, which is part of the reason caution is required when taking this sort of tack, in this case the description of the neurological mechanisms (and their failings) involved is more explanatory than any description from a higher-level. It gives us a way of understanding why there seems to be this set of individuals who engage in similarly bizarre behaviours – the reason is that, in these individuals, there is a similar malfunction in a mechanism we all have.

Similarly, there are many cases of people with Parkinson's Disease who, when put on medication to treat the disorder, develop gambling addictions or other similar issues.[59] The issue seems to be that the extra dopamine is causing an issue with a neurobiological mechanism – the flood of dopamine, which plays a vital role in learning and reward-seeking behaviour, causes a misfiring, causing far more weight to be put on the returns (or potential returns) from gambling than they (normatively) should. It is interesting to note that dopamine is implicated in many behavioural pathologies, including drug abuse.[60] In most of us, the release of dopamine is closely controlled by a group of neurons (dopaminergic neurons).[61] When the control of dopamine is lost through the supplementation of dopamine in medication, there forms a disconnect between the mechanism which controls this release and the groups of neurons which use this release to learn. Dopamine plays a crucial role in learning and motivation, so this disconnect means one of the major controls over behaviour has been lost. Instead of the dopaminergic neurons playing a role in controlling behaviour, they are (to a lesser or greater extent) made less significant, meaning a part of the agent which used to have control over aspects of behaviour no longer does (or, at least, that control is diminished or compromised). This means there is a disconnect at a very deep level, and the fact that this often leads to a seemingly non-normative change in behaviour suggests that this loss of control causes a disconnect between the

[59] Arias-Carrión & Pöppel, 2007
[60] Gaetano, 1995
[61] Arias-Carrión & Pöppel, 2007

agent's true values and their current behaviour. The reason for the behaviour is not from the person's deep values or personality traits. Instead, it is from a mechanism not doing what it should, causing a disconnect between these deep values and the behavioural outcomes.

As noted above, we must be careful not to become mixed up and allow for just anyone to plead that they are not responsible for some action because some brain mechanism made them do it. We cannot define the normal working parameters for brain mechanisms to be those such that the cause of any morally wrong (or right) action falls outside of those parameters, as this would eliminate responsibility – anyone who did something which was wrong or right would by definition would have acted due to a faulty mechanism, and therefore would not be responsible. We can try to identify some cases where it would be wrong to assign responsibility, such as when the action derives from a part of the brain which has a lesion which obviously causes that action or when medication is causing a drastic change in neurochemistry, which has a documented effect on behaviour of certain sorts. These sorts of faulty mechanisms, which lead to a lack of integration across neural mechanisms, may map on to a parallel lack of integration at the psychological level.

There are all sorts of cases which are not so obvious, or where it seems the physiological facts may somewhat mitigate, but not completely eliminate, responsibility. Many addiction cases may fall in this range – substance addictions certainly have a physiological component, and may have an effect on neural mechanisms. There are addictions of varying degrees of severity, and the considerations here help to elucidate the significance of these addictions for attributions of responsibility.

Conclusion

In this chapter, I have attempted to address some of the issues with the Whole Self account by drawing on the Deep Self account. This gives us some criteria for what sorts of causes of actions we can hold an agent responsible for, although I have argued that we may not be able to give a rigorous definition of exactly what these criteria are. I have also made some suggestions about some of the cases

in which responsibility should not be assigned to an agent – if the action stems from a broken mechanism, the agent is not responsible.

Chapter Three

Introduction

The previous chapters have attempted to build up a theory of moral responsibility, by combining and building upon the Deep-Self and Whole-Self accounts. The discussion thus far has offered a general view of the theory. However, this general level view does leaves many important questions unanswered. This chapter will be an attempt to address some of those questions.

The discussion thus far has left in the background several concepts which should be investigated explicitly. In order to do this, I will be drawing upon the empirical sciences. By using the empirical sciences as a basis for expounding upon these concepts, we avoid the possibility of letting our intuitions run unconstrained. By using empirical work to constrain our theorizing, we make sure our theories fit the real world.

Throughout the discussion of the previous chapters, I have frequently brought up issues of over-powering desires – e.g. the compulsions that sufferers of Obsessive Compulsive Disorder and addictions face. Agents with addictions or pathologies are often exempt from being held responsible for actions stemming from these afflictions. This raises a host of issues, which thus far have not been explored – what is will-power, which can be used to combat strong desires? Why, in the cases of OCD and addictions, can agents be excused from responsibility for acting on strong desires? We all face desires which are external to our selves in some sense, but we are expected to be able to exercise enough self-control to keep from acting on such desires. In many of these cases, a failure to control ourselves seems like something we could be responsible for – saying that we acted with a weak-will certainly does not always get one off the hook. What differentiates these cases from the excusing cases? I will argue that we can differentiate compulsions from strong desires in a morally relevant way, and that understanding more about self-control can help to sharpen the distinction.

Bound up with issues of compulsions and self-control are issues of judgement. Generally, we

say that people are acting weak-willed (synonymously, without self-control) if they are acting in a way contrary to what they judge to be in their long-term interests. Part of understanding self-control, then, is understanding judgement and reasoning. Many other questions about judgement have been thus far left untouched, in particular: What is the connection between judgement and responsibility? Our judgement often leads to actions (or lack of action), and it is for actions that we hold people responsible. This is an area unexplored in our theory thus far, yet quite ripe for discussion. I will argue that there are a host of reasons why we may, in some cases, blame people for bad judgement, and that the Whole/Deep Self account can account for this.

The purpose of this chapter is to shed some light on these concepts which have been playing in the background thus far, and to investigate their connections with responsibility. Doing so will strengthen the account given so far, and extend it to deal with some interesting questions about responsibility.

Clarifying Concepts

If anything is capable of having moral responsibility, it seems likely that humans (generally) are. It seems what makes us likely candidates is the fact that we can understand our environment, make judgements about our long-term goals, assess reasons for acting, form intentions on the basis of these assessments, and carry out these intentions. Much of the discussion of moral responsibility has proceeded as philosophy discussions often do: from anecdotes, thought experiments, and appeals to intuition. These tools are put to use exploring the concept of moral responsibility, what it means, and in what situations agents have it. While these tools are, of course, invaluable to the philosopher, often they are being used in conjunction with numerous concepts which are simply assumed, or only vaguely defined. With a topic such as moral responsibility, many terms about agents and minds (such as judgement, rationality, desires, self, etc.) are used without rigorous discussion – with good reason, as many of these concepts are vast topics on their own. However, it may be fruitful to take a closer look at

these concepts, drawing on the empirical sciences for guidance. While we all (presumably) have minds, and therefore have first-hand experience with them, the cognitive sciences have proven to be extremely useful in helping us further our understanding of minds. As Paul Thagard notes, "Philosophy operates best [...] with empirically informed reflection on a wide range of findings in cognitive science."[62] Our concepts and intuitions are sometimes incorrect, and need constraints and guidance from empirical findings.

Minds are very complex, and the properties of these complex entities have ramifications for responsibility. How we perceive the mind as functioning is going to have repercussions on our conception of responsibility. If it is thought that there is no such thing as practical irrationality – if we think that people always act in the way they judge as being best, and always judge rationally – our ideas about attributions of responsibility are going to be quite different than if we see humans as often acting irrationally. Often, if we think someone acted due to poor judgement, or due to a strong temptation, responsibility can be mitigated or – in extreme cases – even removed. Our views on responsibility hinge in important ways on our views of the mind.

While we may believe we have a grasp on the basics of how a mind works, and that the empirical sciences are simply the messy details, we are often wrong about how things work. For example, introspectively it seems that our memories are simply records of past events which can fade or become more difficult to recall. In fact, memories are malleable, and are often filled with incorrect details even though we feel we vividly recall them, as evinced by how readily false memories can be constructed or manipulated.[63] An investigation which drew on the concept of memory, taking introspective certainty to always indicate an accurate record of past events, would be starting from false principles. There are plenty of other examples – for instance, our introspections about our reasons for acting are often wrong.[64] Even though it may seem crystal clear to us introspectively why we are acting,

[62] Thagard, 2009
[63] Paterson et al., 2009
[64] Nisbett & Wilson, 1977

the actual reasons are often hidden from us and our explanations are sometimes just rationalizations. We must acknowledge that often, our intuitive notions about certain concepts are wrong, and we must inform our notions with empirical insights.

There are numerous concepts relevant to us which could be analyzed in this fashion. Terms such as “agent”, “reasons”, “responsibility”, “self”, etc. all could be investigated in great detail to shed light on the concept of moral responsibility. Here, the discussion will be restricted to two interconnected concepts which seem most closely related to the investigation of moral responsibility carried out in the previous chapters: self-control and judgement.

Self-Control

It is interesting to note that sometimes we act for good reasons, sometimes we do not. A great amount of philosophical literature has been written on the subject of weakness of will, and there are a diversity of opinions on what exactly weakness of will is. Richard Holton, for example, holds that weakness of will is the unwarranted changing of intentions.[65] In this view, once one has judged that a certain course of action is best, and an intention is formed to act in that way, revising that intention without a good reason to do so is an example of weakness of will. The other common view is that weakness of will is akrasia – acting contrary to one's judgement. In this view, held by Alfred Mele,[66] acting weakly is acting in one way while at the same time judging that it is not the best course of action. While this is an interesting debate, we do not have time to go into it here. Instead, we will use a general definition of weakness of will which satisfies either: weakness of will is acting in a manner one has judged (either at the time of action or previously) not to be best, without warrant to overturn that judgement.

A distinction must be made here to avoid confusion. Weakness of will is acting contrary to a best judgement, not necessarily what is best objectively. Of course, agents can judge a course of action

[65] Holton, 1999
[66] Mele, 2010

to be the one they have most reason to do, but be incorrect – due to an error in reason, incorrect or incomplete information, etc. Acting contrary to such judgements can still be a case of weakness of will. However, it is likely safe to say that in many cases, a failure of self-control leads to actions we do not have good reasons to do.[67]

Self-control is that which allows us to stick by our intentions for future actions and overcome weakness of will, though precisely how it does this is somewhat of a mystery. Whatever self-control is, it is of central importance to many of the cases we have been discussing in previous chapters. When we withhold blame from someone who has an addiction or Obsessive Compulsive Disorder, we take into account whether or not we feel that they should be able to exercise enough self-control to defeat their urges. Self-control, though, is a rather strange concept – what does it mean to control ourselves? Why should we sometimes find ourselves out of our control, and what does it mean for us to be exhibiting self-control? And why should considerations about self-control have any bearing on responsibility?

The Neuroscience of Self-Control

The ventromedial prefrontal cortex (vmPFC) is a brain region which has been found to play a central role in making decisions. It is thought that this area codes for the value of possible choices.[68] Activity in this area correlates with the value subjects place on the options of a decision they are making. This has been shown to be the case in a number of different decision-making scenarios. This, combined with data that damage to the vmPFC impairs decision making,[69] makes a strong case that the vmPFC is central to the process of attributing values to different courses of action.

However, an interesting issue arises when one considers self-control. Acting weakly often seems to be done against our better judgement – that is, we end up taking a course of action even though we have judged previously that it was not the best way to act. It seems there are two valuations

[67] Though possibly not all of the time – for possible examples of self-control leading to an irrational course of action, see Arpaly, 2000

[68] Padoa-Schioppa, 2011; Rushworth et al., 2011

[69] Padoa-Schioppa, 2011, pp. 333-332

which are made – the self-reflective, conscious admission that what the action was not truly the most highly valued, and the one more directly connected to our motivation which leads to action. When our self-reflective valuation deems one option higher, but our motivational valuation rates the other as more valuable, this leads to a self-control conflict – a conflict between what we consciously know is best, and what we are motivated to do. Does the activity in the vmPFC track our conscious valuation, or the motivational? It seems to be the latter – the choices of dieting subjects correlates with vmPFC activity, not with what their long-term values (supposedly) are, and activity in another brain region, the dorsolateral prefrontal cortex (DLPFC) seems necessary for taking these long-term values into account.[70] It is thought that the DLPFC modulates the valuation process taking place in the vmPFC, making sure it takes into account long-term goals. If this is the case, why does the DLPFC sometimes fail to modulate the valuation in the vmPFC?

Ego Depletion

Though we do not currently know the complete story of why the DLPFC does not modulate the valuation process in the vmPFC, and thus why we sometimes fail at self-control, the phenomenon of ego depletion may give us a hint. Ego depletion is the name given for the phenomenon that certain tasks are able to deplete some resource that seems required for engaging in many types of high-order planning and reasoning, including self-control behaviour. Ego-depleted subjects thus are impaired when it comes to performing self-control tasks.[71] Ego depletion is caused by, and affects, a large number of different tasks. It is at the same time interesting and strange. Engaging in some seemingly unrelated task – such as making a difficult decision – can cause ego-depletion, leading a subject to spend much less time attempting to solve an insoluble puzzle. Ego-depletion leads to lowered performance on many self-control tasks, such as squeezing a hand-grip to exhaustion, or eating radishes

[70] Hare et al., 2009
[71] Hagger et al., 2010, p 495

and avoiding chocolate.[72] It can be caused by asking subjects to suppress their emotions while watching an emotional video, or asking them to perform the Stroop task.[73] Levy points out how general these tasks are:

> The broad and systematic effects of ego depletion are not preferentially exhibited in the domain of self-control or the maintenance of resolutions at all; nor are the effects produced by temptation alone.[..] Ego depletion is produced by Stroop tasks, in which subjects have to name the colors rather than reading the (conflicting) words; Stroop tasks involve the inhibition of responses, but not temptation to break a resolution. It is also produced by having to make choices, which does not involve the inhibition of a response at all. It is even produced by exaggeration of prepotent responses.[74]

Since a wide range of different tasks are both caused by, and cause, ego-depletion, it is difficult to account for these results without making recourse to some sort of shared resource. Subjects who engage in activities requiring this resource (such as suppressing their emotions during an emotional video) go on to show less self-control in seemingly unrelated follow-up tasks requiring self-control (such as squeezing a hand-grip for as long as possible) presumably because the first task saps some of this resource.

The reason, then, for why we do not always act in self-controlled ways, may be that the neurological mechanisms needed to give the proper weight to long-term goals (as well as other forms of high-level reasoning) are very resource intensive. In particular, it seems that they may require glucose,[75] and it may be that we do not constantly engage in self-controlled behaviour because of a biological strategy to conserve the brain's reserves of glucose.

Intriguingly, it has been observed that those who, for at least several weeks, engage in self-control exercises – such as monitoring and improving posture, regulating mood, avoiding sweets, or squeezing a handgrip for as long as possible – then seem to have increased stamina when it comes to

---

72 Hagger et al., 2010

73 In the Stroop task, subjects are simply asked to read the words that appear before them. However, the words are all the names of colours, and are printed in a colour other than the one the word is naming. So, for example, the word “red” may appear, but be coloured blue. This is a surprisingly difficult thing to do, and reaction times are much lower for this task than if the colour of the words matches the colour the word names.

74 Levy, 2011, p 147

75 Gailliot & Baumeister, 2007

other self-control tasks in the lab,[76] or even when it comes to a real-world temptation such as smoking.[77] Subjects who have received self-control training are able to engage in a self-control task for longer than those who do not, and smokers are less likely to relapse after receiving training. Exactly what this training does is not known – it may increase the amount of glucose the brain stores for these resource intensive processes, it may make the processes more energy efficient, or perhaps it changes the resource allocation strategy within the brain, making it more willing to expend the extra resources to engage in self-controlled behaviour.

This section has aimed to provide some understanding of what sort of thing self-control is, and why we sometimes fail to engage in it. By elucidating this concept, we're now better able to approach the issues surrounding the connection between self-control and responsibility.

Self-Control and Responsibility

In many cases, we blame people for their lack of self-control. People who cheat on their lovers, and then attempt to claim that they "couldn't help it", do not evoke our sympathy. We blame people if, through laziness, they do not fulfill their professional or personal obligations. This is generally not mitigated if they afterwards tell us that they knew fulfilling their obligations was the best thing to do, but they could not bring force themselves to do them. Yet, in many other cases, we do not blame people for failing at self-control. People in the grips of a strong addiction, for example, are generally seen as needing help and not to blame for continuing their substance abuse (though perhaps they are to blame for becoming addicted in the first place). How do we separate these cases? In previous chapters, I have argued that certain forms of pathologies which give rise to strong desires are not things we can be held responsible for. This is because giving into them does not properly reflect any aspect of the Deep Self. One might object to the view on the grounds that, in some cases, we do hold people responsible for giving in to temptations. If we only blame agents for actions stemming from their Deep Selves, and

[76] e.g. Muraven et al. 1999
[77] Murvaven, 2010

many temptations are external to the agent's Deep Self, how can we blame such an agent for acting on these temptations? It seems we should not blame any agent for giving in to temptation, yet there are some cases of giving in to desire we blame people for, and others which we do not. Without a way to differentiate these cases, we are left either claiming that people are responsible for compulsions in the same way they are responsible for other strong desires, or that people are not responsible for any strong desires at all.

How, then, should we attempt to differentiate these cases? Watson gives a useful criterion:

> The weak and the strong may be subject to desires of exactly the same strength. What makes the former weak is that they give in to desires which the possession of the normal degree of self-control would enable them to resist. In contrast, compulsive desires are such that the normal capacities of resistance are or would be insufficient to enable the agent to resist. This fact about compulsive desires is what gives substance to the claim that they are too strong.[78]

This gives us a way to differentiate the compulsive cases (cases in which we do not want to hold an agent responsible) from the weak cases (cases in which we do). If the temptation which causes the agent to act is one that requires more than the normal capacities for self-control, then we should not hold them responsible. This makes sense – knowing that someone cannot overcome a temptation which most people would not be able to does not tell us much about that person's deep-self.

However, this raises the question: What are the normal capacities for self-control? Watson can do little more than gesture at this. However, with empirical tools on our side, we can be a bit more descriptive. We have noted that engaging in self-control seems to deplete some resource, and that as this resource becomes depleted, we become more likely to act weakly. We have also noted that people can become better at self-control, and there is empirical support for individual differences when it comes to self-control. These differences may come from the way the resources for self-control are used: individuals may vary in the amount of the resource stored, or with the efficiency with which they use it, or with the conservation strategies deployed in using that resource. This gives us an idea of the mechanism for the variation of strengths of wills. We could run empirical studies to measure what the

[78] Watson, 1977, p. 330

range of people with normal-strength wills would do in various situations to come up with what constitutes the normal capacities for self-control which Watson makes reference to. It is important to note that although we do not have the results of such empirical studies to rely on when making judgements about responsibility, we likely approximate. We likely think about whether we would be able to resist a temptation in a similar circumstance, or whether our experience tells us that most people in similar circumstances would be able to. This gives us the ability to judge whether a particular case was an example of someone acting weakly, or being compelled.

Why do we hold people responsible for their lack of self-control in the non-compulsive cases? The Deep/Whole-Self view has a couple of possible answers, each of which may contribute somewhat to the intuition that sometimes people are to blame for their weak wills. One possibility is just that failure to engage in self-control may actually come from the agent's apathy. That is, it could be that the agent judges that taking a particular action does lead to better long-term outcomes, but judges that this is only slightly preferable to the short-term gains of doing otherwise, and thus is not particularly motivated to engage in self-control. This may, in fact, reveal deep preferences and values by showing that these future prospects do not mean much to the agent, thus giving warrant to any blame that may be placed. Another possibility is that we consider weakness as a deep trait. Since compulsions only show that a person does not possess greatly above-average strength of will, these do not reveal any such trait – it only shows that they fall somewhere in the normal range in their will power, something which presumably most people do. But if one acts weakly (especially if done so consistently), this may reveal that they are a weak-willed person, the sort of person who is just unable to resist temptations. Being unable to resist temptations may be a negative trait just as being selfish or having other forms of ill-will are.

We have addressed the concept of self-control, what it seems to be, why it has some of the properties it does, and what its implications for responsibility are. The neurological data on self-control points to the DLPFC playing a role in mediating the valuation process in the vmPFC, which allows for

long-term interests to be taken into consideration. The phenomena of ego-depletion suggests that this neural mechanism is resource intensive, which is why we do not always engage in self-control. The distinction between compulsion and weakness helps to differentiate cases in which an agent may be held responsible for their actions versus when they may not. Acting weakly may reveal something deep about an agent by showing deep biases within the agent, or by showing apathy about the issue at hand. A lack of strength of will itself may be a deep trait which we hold individuals somewhat responsible for. Thus, the Whole/Deep Self can account for why we, in some cases, blame people for acting weakly.

Judgement

Judgement plays an important role in both self-control and in issues of responsibility. Weakness of will is often described as acting contrary to one's best judgement, and self-control is supposed to stop weakness of will, helping us act in ways consistent with what we judge is best. Ergo, judgement and self-control are quite closely linked. We hold people responsible for actions – according to the Whole/Deep-Self view, actions which stem from deep values and desires within the agent. Acting, however, always (unless it is a reflex or in some other way unintentional) requires a judgement about how to act, whether it is a quick, unconscious judgement, or the result of a long, conscious deliberation.

As noted above, a region of the brain – the vmPFC – seems to play a large part in assigning values to different options of choices we face. However, the values it comes to do not always properly weight our long-term interests – modulation by another brain region, the DLPFC, is necessary, and this does not always happen. Even if, upon reflection, we realize our choice is not the best, we may end up choosing the inferior option in any case. This is weakness of will, and it stems from two different valuation processes – our conscious deliberation about what is best, and the process in the vmPFC, which is more closely tied to motivation.

Presumably, though, these two valuation processes are connected. It is unclear how much conscious thought affects the vmPFC's valuation process as it is going on, but the fact that it can take

into account long-term goals at all shows that our higher-order planning and goal-setting can have an affect on how we value different courses of action. Of course, this is not surprising – we know that we can, at least sometimes, act with strong wills, facing short-term pain for greater long-term rewards. The last section was about how we do this, and why sometimes we fail to. This section is about judgement and reasoning – specifically, how we come to consciously judge particular goals as worth pursuing.

Much of the psychology research in reasoning and decision making works within the dual systems theory framework. The basic idea behind dual systems theory is simple: there are two systems involved in reasoning: System 1, which is quick, heuristic based, and prone to bias;[79] and System 2, which is slower, more deliberative, can correct System 1, and can involve abstract reasoning and hypothetical thinking.[80] It is assumed that System 1 is evolutionarily old, and that System 2 is newer, responsible for higher-level reasoning, and is part of what makes us unique as human beings. There is evidence that strongly suggests that System 2 reasoning requires more physiological resources – particularly glucose – and that when these resources are low, we are less likely to engage in System 2 reasoning.[81] In fact, this is the same resource required for self-control, making this an example of ego-depletion which we discussed above.[82] Although this account may over-simplify, positing only two systems, there are robust findings which the dual systems theory makes sense out of. Even if the story is more complicated than the dual systems view would have us believe, it seems the story may at least be approximately correct.

The dual systems account of reasoning is often used in the decision making literature to account for many sorts of seeming irrationality. Typically, it is assumed that when subjects in a judgement task make their choice based on irrelevant details, or fail to use relevant data such as base-rates, the heuristics-based System 1 reasoning is in use; when these biases are overcome, or reasoning is using

[79] Morewedge & Kahneman, 2010
[80] Evans, 2003
[81] Masicampo & Baumeister, 2008
[82] Levy, 2011

complex data such as base-rates, it is assumed System 2 is active.[83]

Similar to how we do not always engage in self-controlled behaviour, we do not always engage in System 2 reasoning. Researchers of decision making are interested in what activates System 2 reasoning, and have made some headway discovering what does this. For example, it seems that some amount of metacognitive difficulty may be what activates System 2 reasoning, at least in some circumstances.[84] When we run up against something that we are having difficulty processing, we call in the metaphorical big guns – System 2 reasoning. If no such difficulty is encountered, we go on with our quick, low-effort, heuristics based approach.

Though this does not explain exactly what judgement is, this discussion does help to clarify the concept somewhat. Judgement and reasoning are complex cognitive functions. When we engage in reasoning, sometimes we use one set of cognitive tools based on heuristics which usually guide us well (System 1), and sometimes we use a different set which, though they take more effort and resources, allow us to engage in more abstract reasoning which is less subject to bias (System 2). With this additional understanding of judgement, we are now able to ask an interesting question about responsibility: To what degree are agents responsible for poor judgement?

Judgement and Blame

Judgement is an important part of moral responsibility for a fairly obvious reason: We use our judgement to choose our actions, and what we hold people responsible for are those actions that they choose (in some sense). I want, then, to elaborate on this connection in two ways: first, to be explicit about the role judgement plays in the Whole/Deep-Self view outlined in previous chapters. Second, I want to look at casting blame at an agent for their poor judgement.

The first connection is fairly simple. The Whole/Deep-Self view holds that an agent is responsible for an action if that action stems from a desire or value (or lack thereof) which constitutes a

[83] e.g. Alter et al. 2007
[84] Alter et al. 2007

deep part of their self. However, desires and values do not spontaneously generate actions we should take. When we are faced with a decision, we use our judgement to decide what to do. Our judgement connects our desires and values with outcomes of various actions, and attempts to come up with the outcome which best serves our interests. Judgement can be good or bad – that is, it can optimally (or near-optimally) choose an option which is best suited to satisfy our desires, or fail in some way to come up with an optimal solution (by improper weighting of desires, poorly approximating the likelihood of some event, or failing to take some factor into account at all).[85] This basis for evaluating judgements is internalist – it considers whether the judgement is good or not in terms of the agent's internal perspective, based on the agent's interests. A discussion of an externalist pespective, which considers broader, moral standards, will occur in the next section.

Given this first connection, the second becomes even more interesting: if bad judgement is possible, we are responsible for those actions which stem from our values and desires, and bad judgement implies somehow failing to properly connect our values and desires to the option we choose, are we responsible for actions we take through bad judgement? There is an appeal to answering this question in the negative. Since poor judgement seems to just be a mistake on the part of the agent, it is hard to see why we may want to hold people responsible for their poor judgements. However, I will argue that the answer to the above question is a qualified “Yes”. The Whole/Deep Self view posits that we are responsible for those actions which stem from our Deep Selves. If poor judgement causes a disconnect between our actions and our deep values, will the Whole/Deep Self view be able to account for cases in which we attribute blame to agents taking action through poor judgement? The answer to this question is, again, a qualified “Yes”.

The answer to the first question relies on intuition, but I believe it is a common intuition. When a politician (or anyone in a position of sort of power) makes a mistake despite having a vested interest

[85] Arpaly (2000) argues that there are cases in which our best judgements can fail to track what we actually have most reason to do (from an internalist perspective), even though an unconscious judgement may be tracking the reasons much more faithfully.

in doing a good job, it feels right to blame them for this mistake. Imagine a politician who, through honest but poor economic management, managed to plunge his/her country into a financial disaster. The politician does want to be re-elected, and cares for the state of the country. However, through a series of blunders, the politician has brought the country to financial ruin. It feels that at least to some extent, this politician is blameworthy. Certainly, this blame is different than if the politician acted out of selfish reasons (if, for example, they had been paid off by a rival country to do this), but it still seems there is some amount of blame to be placed, despite being an “honest mistake”. Even though the consequences of the politician's decisions did not stem from the politician's deep values, it still seems that blame is warranted, at least somewhat. Something being an honest mistake may alleviate some of the blame, but not all of it, and a theory about responsibility should be able to tell us why we are holding people responsible in these cases.

To elaborate on the answer to the second question, we must return to the discussion above about dual systems reasoning. Recall that System 1 is quick, heuristics-based, takes less effort, and is more prone to biasing, while System 2 is slower, required to engage in higher-level reasoning, and works to correct System 1. So we are more likely to make mistakes in reasoning when engaging in System 1 reasoning, and engaging in System 2 reasoning is costly. This leads to a fairly intuitive conclusion – putting extra effort into judging an issue makes one less likely to make mistakes. This leads to one way in which the Whole/Deep-Self account can explain holding people accountable for poor judgement – similar to the above argument about self-control and apathy, making mistakes in reasoning may reveal something deep about the person, because it may show the agent does not care enough about the issue to put in the extra effort. In this case, the results of the judgement, or the resulting action itself, may not reveal much about the agent's Deep Self, but if it is clear to outsiders that the reasoning itself seemed to be poor, this may give pretty strong evidence about the agent's deep values – it may show, even in cases in which an honest mistake has been made, that the agent does not place a high value on the outcome of this judgement.

Sometimes, it may not be that we blame agents just for making errors in judgement, but that we blame them for the type of errors. Biases often influence our reasoning, especially when engaging in System 1 reasoning. Some of these cam be morally innocuous (such as hindsight bias, where one sees a past event as more predictable than it was – and recalling having predicted it to greater accuracy than one actually did). However, others may show something deep about an agent. A judge who mistakenly declares a suspect guilty may be suspected of racism if the mistake was large enough, the suspect an ethnic minority, the judge an ethnic majority, and no other factors seem to explain the judge's mistake. Not all cases may be as straight-forward as this, but we often implicitly suspect biases of this sort. For example, we may often implicitly suspect politicians or pundits of unconscious biases (seeing the poor, rich, foreigners, or members of the opposing party in a particular way) and this may play a role in our condemnation of those people. In this way, a mistake can reveal a deep rooted value in the form of biases. This sort of bias has been investigated empirically, for example, in showing that racial biases can affect legal decisions.[86] While we all have biases, and cannot be blamed for being imperfect humans, a strong bias can certainly reveal something important about a person's deep values.

Deep values may also be revealed in a poor judgement in a more direct way. Even if an agent's judgement is poor in the sense that, according to their deep values, they should have chosen another option, this does not mean that their choice does not stem from deep values. Consider, again, a politician, making serious policy decisions, who is stuck with one of two options: 1) choose a policy which is best for the country, but is unpopular with voters, or 2) choose a policy which is popular with voters, but damaging to the country. The politician does deeply care about the country, but also cares deeply about being re-elected. If a proper judgement was made, the politician would choose the unpopular option, as the value placed on the well-being of the country is, subjectively, higher than the value the politician places on being re-elected. However, at the end of a long day, the politician is forced to make a choice – and makes a poor choice, according to her/his deep values, and chooses the

[86] Levinson, 2006

option which is worse for the country. This choice was, largely, because of the agent's deep desires – the politician really does want to be re-elected. Even though, if able to give it more thought, the politician would make a different choice, the choice made still does reveal some deep values. Thus, the outcomes of of some bad judgements may explicitly and consciously be caused by deep values.

I have outlined three ways in which a mistake in judgement may reveal something about the Deep Self. This, however, is unlikely to tell the full story. There may be other ways in which mistakes in reasoning reveal something about one's deep values and desires, but more importantly, we may also simply see one's ability to judge as an important aspect of them, which we can hold them responsible for. When someone is being blamed for a mistake in judgement they have made, it is not uncommon to hear them condemned as “stupid”. Calling one stupid can be used as more than just a derogatory term – it can be a term of condemnation. Perhaps, then, we simply see agents' faculties of judgement as an integral enough part of who they are to blame them for having a poor one. This may seem unfair, since some “luck out” and have good judgement, some get the short end of the stick and have poor judgement. However, this is also the case when it comes to good or ill will – some people may be born with, or are raised to have, selfish attitudes. Yet we blame people for selfish attitudes, and to a much, much greater degree than we blame people for honest mistakes (even without correcting for unconscious biases or apathy).

It seems that in some cases, we do blame people for their poor judgements. While saying something is an “honest mistake” may, if truthful, mitigate some of the blame, it does not always eliminate it. I have suggested a few different ways of accounting for this, in the framework of the Whole/Deep Self view. The first three ways posit that a poor judgement actually does reveal something about an agent's deep values, and therefore (on the Whole/Deep Self account) the agent can be held responsible for them. Poor judgements may show low value being placed on the subject matter being judged, revealing the agent's deep values (or lack thereof) about the subject matter; it may be caused by deep, unconscious biases; or may be more explicitly caused by conscious, deep desires which are just

being weighted incorrectly. Finally, I also suggested that perhaps we also place some blame on people who have bad judgement, just for their having poor judgement – we may treat judgement itself as a part of the Deep Self.

Judgement and Pathologies

Some people simply have poor judgement. They are just regularly unable to judge things well. However, an agent may be said to have poor judgement in one of two ways. Above, the discussion centered around judges which are bad from an internalist perspective – what constitutes a poor judgement is what does not lead to the agent's best interests. Judgement which is poor from an internalist perspective does not satisfy the agent's desires (short term and long term) in a manner that would be optimal. Here, I want to consider those with chronically poor judgement in another sense: being poor judgement from an external, morally normative perspective. If an agent's judgement regularly does not take into account factors that it (normatively) should, there is a sense that we can say that the agent's judgement is poor, even if it is serving the agent's best interests.

According to Wolf, for an agent to be responsible, more is needed than reason-responsiveness and for actions to stem from the deep-self. She argues that many views of responsibility and the self (she focuses specifically on Frankfurtian True Self views, but her argument applies just as well to the view defended here) miss a crucial condition: there is a sanity condition on responsibility.[87] This means one's self and values be connected to the world in a certain way. On Wolf's view of moral responsibility, people who grow up with deprived childhoods or in misguided societies have their actions governed by mistaken conceptions of values, and are therefore not responsible for actions stemming from their mistaken values.

For Wolf, agents who grow up in an environment which shapes them such that they cannot tell right from wrong are not to blame. If we imagine a ruthless, sadistic dictator who has given an education to her son which insures that he will grow up to hold the same values as his mother, Wolf

[87] Wolf, 1988

says the resulting man is not morally responsible for acting on these values later in life – even if he fully embraces who he is and the values he holds. The reason for this is that such an agent's values are not rooted in the world. The agent cannot be said to value what he does because he understands what is good about these values – as Wolf says about racism, "We cannot say that the racist is responsible for his racism if it results from his understanding of what is good about racism – for there is nothing good about racism for him to understand."[88] This creates an asymmetry between agents who are raised with good values and those who are raised with negative ones, since the agent who is brought up with positive values can recognize and accept what is good about their values, while there is no such option for the agent brought up with negative values. For Wolf, this means that the agent brought up with good values can be held responsible, while the agent brought up to be a racist or a ruthless dictator cannot.

This argument can be understood in terms of the reasons-responsiveness criteria of Fischer and Ravizza[89] outlined in the first chapter. Recall that, according to Fischer and Ravizza, moderate reasons-responsiveness is required for moral responsibility. Moderate reasons-responsiveness includes strong reasons receptivity, which means not only being able to recognize reasons for acting, but your hierarchical ordering of reasons must be understandably patterned in an objective way. If you would pay $200 for a ticket to the basketball game, you should still want to go if you find out the ticket is only $100. Wolf's argument, then, can be understood as saying that the racist (or anyone else brought up with sufficiently bad values) is not properly reason responsive – if they accept the value of racism, they should be even more willing to place value on equality, since it is objectively better. If the racist is confronted with the value of equality, but rejects it in favour of racism, the racist has failed to rank these in an objectively rational way, is not strongly reason receptive, and hence cannot be held responsible.

There are two assumptions in this analysis which need to be made explicit. The first, most obvious one is a particular view of morality – that morality is, to a large extent, objective. There is

[88] Wolf, 2005, page 269
[89] Fischer & Favizza, 1998

room for disagreement with such a view. This is a very large issue on the nature of morality, which can be put aside here – suffice to say, if one does not share Wolf's view on this matter, one has sufficient reason to disagree with her conclusion that agents with a poor moral upbringing cannot be responsible. I will remain agnostic on this issue, as the view outlined thus far does not hinge in any way on it. The second assumption made by the above analysis is that there is no sufficient non-moral reason for preferring an inferior moral value (such as racism). One possibility is that the racist, after being confronted with the ideals of equality, keeps with his old, racist ideals because of the effort it would take to adopt new ideals, the social pressure to keep the racist ideals, the feeling of superiority which one gains from seeing those of a different ethnicity as inferior, etc. The racist may, at some level, all else being equal, value equality over racism morally, but all else may not be equal, and there may be other reasons for continuing to hold racist ideals. In that case, the racist may be strongly reason receptive, and a proper target of blame.

Thus, depending in important ways on one's views of the nature of morality and on the specific story one tells about why an agent does not accept a better moral value (i.e. whether they do not because they are incapable of recognizing it as better, or whether they do not because the agent has some non-moral reasons for choosing the inferior moral option), one may or may not agree that those raised with poor moral values are not to be considered reason receptive. Thus, there is certainly room, in the Whole/Deep Self account (which posits reasons-responsiveness as a prerequisite for blame), for one to say certain people raised without proper moral training are not blameworthy. But this conclusion is not a necessary one.

The above-mentioned racist and the dictator's son are examples of agents who have been shaped by their environment to have particular, warped values. But there are also cases of pathologies which leave certain values inaccessible to the agent. Psychopaths are are an interesting case because they are unable to make a distinction between moral norms and social norms.[90] They also have no problem

[90] Blair, 1994

violating either of these norms if it benefits them. They are, in a sense, amoral beings – morality does not factor into their decisions.

Given certain assumptions, psychopaths can be said to be reasons receptive – they can understand reasons, and simply do not consider moral or social norms very good reasons for action. However, just like with the racist, we may question whether psychopaths really do have reasons receptivity, and in the same way. Given the assumption that, all else being equal, it is objectively better to not cause harm than to cause it, psychopaths would fail to be reasons-responsive – if they are truly amoral, they should not care one way or another about harm. Unlike with the racist, there is no non-moral reason which favours this attitude, as there is no social pressure to be amoral, and psychopaths don't derive pleasure from breaking moral norms (they just don't mind doing it). So, given an assumption about the objectivity of some (any) moral valuation, and given that psychopaths do not care one way or another about morality, they are objectively wrong in their hierarchical ordering of reasons, and therefore not strongly reasons receptive. It may, then, be more appropriate to put psychopaths in the same class as bears and wolves and other non-moral beings which can cause harm.

Depending on one's views of the nature of morality and people's motives for holding certain values, people with poor moral judgement may or may not be held responsible. If any moral value holds objective worth, then it seems psychopaths cannot be said to be reasons-responsive, and are not proper targets for blame. Agents who were raised with distorted values may not be proper targets for blame if their values are objectively wrong and there is no non-moral reason for them to hold that value. Thus, there is room for a plurality of views about the blameworthiness of various cases within the Whole/Deep Self view.

Conclusion

The previous two chapters developed a theory of moral responsibility. However, it rested on

many concepts which were left undefined and in the background. This chapter has been about bringing them to the foreground, and drawing connections between them and the issues raised in previous chapters. By drawing on empirical findings, we are able to generate a much richer conceptual framework to work within, and keep our speculations in check. We have drawn on the cognitive sciences for information about judgement and self-control, and learned that both self-control and reasoning require some depletable resource. We often hold people responsible for acting weakly, and this may be because their behaviour shows an unwillingness to expend the effort to engage in self-control, or a weak-will itself may be a deep trait we blame people for. We may hold people responsible for their (from an internalist perspective) poor judgement partially for the same reasons (that judgement is a deep trait and failing to engage in careful judgement may show an unwillingness to expend the effort), but also because their poor judgements may be caused by unconscious biases or other deep values. People who have poor judgement from an externalist perspective may or may not be held responsible, depending on one's views of the nature of morality and on the specifics of the case. These explorations have helped to extend and strengthen the account which was given in earlier chapters.

Summary and Closing Thoughts

The issue of moral responsibility is complex, with many interconnected issues. The central argument here has been that an agent is considered morally responsible only for actions which stem from their deep selves. Addictions and disorders can cause us to act in ways which are incongruent with our deep selves. We do not blame people with Obsessive Compulsive Disorder for their compulsions, because we know in a sense their actions do not stem from themselves – they are caused by an external factor, this disorder, which happens to be able to dominate the motivational system of these people. People are not responsible for all actions that their bodies perform.

I have, along the way, endorsed a Strawsonian view of responsibility[91] – responsibility is an attitude we hold towards other social agents. However, to fill out this view, we must discover under what circumstances this social attitude is justified. The reasons-responsiveness view brings us part of the way. Agents must have an objectively coherent hierarchical ordering of reasons for acting – if an agent will act in one way at one cost, lowering that cost while keeping everything else equal should result in the same action on the part of the agent. There also should be some conceivable set of circumstances which, without any changes to the agent, would give the agent enough reason to act otherwise. This forms a solid foundation for understanding what is required for moral responsibility, but it leaves some very serious questions. There are some cases in which agents seems to satisfy the reason-responsive criteria, yet we would not want to hold them responsible for their actions because some of those actions seem to be caused by something external to them.

In order to deal with this issue, we need an understanding of what sorts of actions stem from the agent. I introduced the Whole Self and Deep Self views to deal with this issue. Agents are responsible for actions which stem from their Deep Selves. Parts of the Deep Self tend to be enduring, desires within it tend to be general (as opposed to directed towards specific things), they tend to be reflectively endorsed by the agent, these traits are often central to the agent's self-conception, and are not generally

[91] P.F. Strawson, 1962

in extreme conflict with other deep traits. This definition excludes pathologies and addictions, and the desires they give rise to are seen as external to the agent's Deep Self.

We often hold people responsible for the actions they take after failing to judge the situation properly, or failing to control themselves in the face of temptation. Agents who act without exercising good judgement or self-control may, in some cases, be doing so for reasons which stem back to the Deep Self, even if the desires for that action itself are not particularly deep.

Moral responsibility is a large issue, which connects with many other deep philosophical questions. The view here provides a framework for thinking about this issue, and furthers the discussion on what moral responsibility is, and when we may justifiably hold an agent responsible.

Bibliography

Acker, F. 2008. “New findings on unconscious versus conscious thought in decision making: additional empirical data and meta-analysis”. *Judgement and Decision Making*, 3:4, pp. 292-303.

Arias-Carrión, Ó. Pöppel, E. 2007. “Dopamine, learning, and reward-seeking behavior.”. *Acta neurobiologiae experimentalis*, 67:4, pp. 481-488.

Alter, A. Oppenheimer, D. Epley, N. Eyre, R. 2007. “Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning”. *Journal of Experimental Psychology*, 136:4, pp 569-576.

Arpaly, N. 2000. “On Acting Rationally Against One's Best Judgment”. *Ethics*, 110, pp. 488-513.

Arpaly, N. 2005. “How it is not “Just Like Diabetes”: Mental Disorders and the Moral Psychologist”. *Philosophical Issues*, 15, pp. 282-298.

Arpaly, N. 2006. *Merit, Meaning, and Human Bondage*, Princeton University Press, New Jersey.

Arpaly, N. Schroeder, T. 1999. “Praise, Blame, and the Whole Self”. *Philosophical Studies*, 93, pp. 161-188.

Blair, R. 1995. “A cognitive developmental approach to morality: investigating the psychopath”. *Cognition*, 57, pp. 1-29.

Damasio, A.R., Tranel, D., and Damasio, H. 1990. “Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli”. *Behav. Brain Res.* 41, pp. 81–94.

Dijksterhuis, A., Bos, M., Nordgren, L., van Baaren, R. 2006. "On making the right choice: The deliberation-without-attention effect". *Science*, 311, pp. 1005–1007.

Dworkin, G. 1970. “Acting Freely”. *Nous*, 4, pp. 367–383.

Evans, J. 2003. “In two minds: dual process accounts of reasoning”. TRENDS in Cognitive Sciences, 7:10, pp 454-459.

Fischer, J. & Ravizza, M. 1998. *Responsibility and Control*. Cambridge University Press.

Fischer, J. & Ravizza, M. 2000. “Replies”. *Philosophy and Phenomonological Research*, 61:2, pp. 467-480.

Frankfurt, H. 1969. “Alternate possibilities and moral responsibility”. *Journal of Philosophy*, 66:23.

Frankfurt, H. 1971. “Freedom of the Will and the Concept of a Person”. *The Journal of Philosophy*, 68:1, pp. 5-20.

Frankfurt, H. 1987. “Identification and Wholeheartedness”. In Schoeman, F. (Ed.), *Responsibility, Character and the Emotions* (pp. 256-286), Cambridge University Press.

Gaetano, D. 1995. “The role of dopamine in drug abuse viewed from the perspective of its role in motivation”. *Drug and Alcohol Dependence*, 38:2, pp. 95-137.

Gailliot, M. Baumeister, R. 2007. “The Physiology of Willpower: Linking Blood Glucose to Self-Control”. *Personality and Social Psychology Review*, 11, pp 303–327.

Graybiel, A. Rauch, S. 2000. “Toward a Neurobiology of Obsessive-Compulsive Disorder”. *Neuron*, 28, pp. 343-347.

Hagger, M. Wood, C. Stiff, C. Chatzisarantis, N. “Ego-Depletion and the Strength Model of Self-Control: A Meta-Analysis”. *Psychological Bulletin*, 136:4, pp 495-525.

Hare, T. Camerer, C. Rangel, A. 2009. “Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System”. *Science*, 324, pp 646-648.

Holton, R. 1999. “Intention and Weakness of Will”. *Journal of Philosophy*, 96, pp. 241-262

Hume, D. 1748. *An Enquiry concerning Human Understanding.*

Kane, R. 1998. *The Significance of Free Will*, Oxford University Press, New York.

Knobe, J. 2003. “Intentional action and side effects in ordinary language”. *Analysis*, 63:3, pp. 190-194.

Levinson, J. 2006. “Forgotten Racial Equality: Implicit Bias, Decision-Making and Misremembering” *bepress Legal Series.* Working Paper 1630. http://law.bepress.com/expresso/eps/1630

Levy, N. 2011. “Resisting 'Weakness of the Will'”. *Philosophy and Phenomenological Research*, 82:1, pp 134-155.

Marino, P. 2009. “On Essentially Conflicting Desires”. *The Philosophical Quarterly*, 59:235, pp. 274-291

Masicampo, E. Baumeister, R. 2008. “Toward a Physiology of Dual-Process Reasoning and Judgement”. *Psychological Science*, 19:3, pp 255-260.

Mele, A. 2000. “Reactive Attitudes, Reactivity, and Omissions”. *Philosophy and Phenomonological Research*, 61:2, pp. 447-452.

Mele, A. 2006. “Fischer and Ravizza on Moral Responsibility”. *The Journal of Ethics*, 10, pp. 283-294.

Mele, A. 2010. “Weakness of Will and Akrasia”. *Philosophical Studies*, 150:3, pp. 391-404

Morewege, C. Kahneman, D. 2010. “Associative processes in intuitive judgement”. TRENDS in Cognitive Sciences, 14:10, pp 435-440.

Muraven, M. Baumeister, R. Tice, D. 1999. “Longitudinal Improvement of Self-Regulation Through Practice: Building Self-Control Strength Through Repeated Exercise”. *The Journal of Social*

Psychology, 139:4, pp 446-457.

Muraven, M. 2010. “Practicing Self-Control Lowers the Risk of Smoking Lapse”. *Psychology of Addictive Behaviors*, 24:3, pp 446-452.

Neely, W. 1974. “Freedom and Desire”. *Philosophical Review,* 83, pp. 32–54.

Nisbett, R., Wilson, T. 1977. "Telling More Than We Can Know: Verbal Reports on Mental processes". *Psychological Review*, 84:3, pp. 231-259.

Padoa-Schioppa, C. 2011. “Neurobiology of Economic Choice: A Good-Based Model”. *Annu. Rev. Neurosci.* 34, pp. 331-357

Paterson, H., Kemp, R., Forgas, J. 2009. "Co-Witnesses, Confederates, and Conformity: Effects of Discussion and Delay on Eyewitness Memory". Psychiatry, Psychology and Law, 16:1, pp. 112-124.

Rushworth, M., Noonan, M., Boorman, E., Walton, M., Behrens, T. 2011. “Frontal Cortex and Reward-Guided Learning and Decision-Making”. *Neuron,* 70, pp. 1054-1069

Russell, P. "Hume on Free Will". *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/hume-freewill/>.

Sripada, C. 2010. "The Deep Self Model and asymmetries in folk judgments about intentional action". *Philos Stud*, 151, pp. 159–176.

Strawson, G. 1986. “On 'Freedom and Resentment'”. In Strawson, G. *Freedom and Belief*, Oxford University Press.

Strawson, P.F. 1962. “Freedom and Resentment”. *Proceedings of the British Academy* 48, pp. 1-25. Rpt. in Fischer, J & Ravizza, M. (Ed.). 1993. *Perspectives on Moral Responsibility* (pp. 45-66). Cornell University Press.

Stump, E. 1993. “Intellect, Will, and the Principle of Alternate Possibilities”. In Fischer, J & Ravizza, M. (Ed.). 1993. *Perspectives on Moral Responsibility* (pp. 211-234). Cornell University Press.

Talbert, M. 2008. “BLAME AND RESPONSIVENESS TO MORAL REASONS: ARE PSYCHOPATHS BLAMEWORTHY?” *Pacific Philosophical Quarterly*, 89, pp. 516-535

Thagard, P. 2009. “Why Cognitive Science Needs Philosophy and Vice Versa”. *Topics in Cognitive Science*, 1, pp. 237-254.

Watson, G. 1977. “Skepticism about Weakness of Will”. *Philosophical Review*, 86:3, pp. 316-339

Watson, W. 1987. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme”. In Schoeman, F. (Ed.), *Responsibility, Character and the Emotions* (pp. 256-286), Cambridge University Press. Rpt. in Fischer, J & Ravizza, M. (Ed.). 1993. *Perspectives on Moral Responsibility* (pp. 119-148). Cornell University Press.

Weatherford, R. 1991. *The Implications of Determinism*. London, Routledge.

Wolf, S. 1988. “Sanity and the Metaphysics of Responsibility”. from Schoeman, F. (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology.* New York: Cambridge University Press, 1988.

Wolf, S. 1990. “The Real Self View (In Which a Nonautunomous Conception of Free Will and Responsibility Is Examined and Criticized)”. In Wolf, S. *Freedom within Reason*, Oxford University Press.

Wolf, S. 2005. “Freedom within Reason”. In Taylor, J. *Personal autonomy: new essays on personal autonomy and its role in contemporary moral philosophy* (pp. 270-286), Cambridge University Press