The Value of Responsibility

András Szigeti

Submitted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy

Date of submission: October 1, 2007

Supervisor: Professor *János Kis*

Acknowledgements

As perhaps usual, this dissertation also took much longer to write than originally planned. It is also for this reason that I have to thank so many people who in various ways helped to make it happen at last. My supervisor at Central European University, Professor *János Kis,* has been a source of inspiration, invaluable criticism as well as a constant example throughout of how to combine intellectual rigour with creativity of philosophical thought which example, I fear, I could only poorly emulate in this work. *Viktor Böhm* has been my friend and intellectual companion from times immemorial and hence the ways in which he has contributed to this project are really just too numerous to recount. *Yehuda Elkana,* President and Rector of CEU, has my warmest gratitude for giving me all the support I needed and my admiration for his unique blend of courage and generosity of heart. I also want to thank Professor *John Finnis,* Professor *John Gardner* and Dr *Martha Klein* for enlightening discussions during my year in Oxford and Professor *Liam Murphy* (NYU), Professor *Arthur Jacobson* (Cardozo), Professor *Michel Rosenfeld* (Cardozo), Professor *Michael Otsuka* (University College London) and Professor *David Wiggins* for the same in New York, London and Oxford.

For comments, criticisms and an ongoing exchange of ideas I am grateful to many members, faculty as well as graduate students, at the Department of Philosophy and the Department of Political Science at Central European University including Professor *Gábor Betegh* and Professor *Katalin Farkas* as well as *Zsolt Novák, Péter Rauschenberger* and *András Simonyi.* I also profitted from the meticulous reading of some of these chapters by *Thomas Donahue* (Johns Hopkins, PhD) and *Matthew Lee* (Notre Dame). An earlier version of Chapter 4 was presented in Oxford at a meeting of the Ockham Society and of Chapter 5 at the Bled Philosophical Conference (Slovenia) on Freewill and Responsibility in 2006. On both occasions, I received important comments and queries.

I thank Central European University for providing an intellectually stimulating and supportive work environment all throughout these years.

Abstract

Contrary to many writings on moral responsibility, this work focuses on the normative implications and justification of responsibility-ascriptions rather than the metaphysical pre-conditions thereof.

The dissertation seeks to answer the following questions: What is the practical significance of ascribing responsibility? What justification can be found for this practical significance? In order to answer these questions, the work adopts the assumption that the necessary and sufficient conditions of morally responsible agency can be met either because compatibilism is true or because libertarianism is true.

The dissertation argues for the Priority Thesis which combines two basic assertions. *First,* that ascriptions of responsibility are based on judgements both in a normative and descriptive sense. Ascriptions of responsibility can guide our behaviour and shape our relationship to other agents by virtue of being judgements. At the descriptive level too, ascriptions of responsibility are motivationally effective because they are taken to be based on judgements. The *second* assertion is that the appropriateness of the judgement of responsibility is necessary for the appropriateness of a distinct range of manifest responses. Being responsible is prior to holding responsible. The Priority Thesis is defended through a critical analysis of alternative theories of responsibility all of which pose a challenge to it. By exposing the shortcomings of these theories a negative argument is mounted in favour of judgement-based account of responsibility.

It is then asked whether a theory of responsibility based on the Priority Thesis can account for the practical significance of responsibility-ascriptions, i.e. explain and justify their normativity. The most common account, the Package Deal Argument, derives the normativity of responsibility-ascriptions from the normativity of moral requirements. This solution is dismissed because it is incapable of explaining why we attach normative significance to the *voluntary* violation (or meeting) of moral requirements.

The Value Thesis is proposed as an alternative solution. The Value Thesis stipulates that ascriptions of responsibility track the value of being a person capable of recognizing and acting on reasons. On this alternative conception, the normativity of responsibility-ascriptions is based on what we value about people. Therefore, responsibility-ascriptions generate reasons for us not only insofar as they trace the violation (or meeting) of moral requirements. A *positive* consideration in favour of the Value Thesis is that we ascribe responsibility for voluntary actions even if those actions do not fall under moral norms, e.g. for the choice of one's life-plan. Finally, it is argued that being a responsible agent is a valuable aspect of personhood because being a responsible agent is *constitutive* of what it is to be a person.

Keywords: freewill, responsibility, morality, moral responsibility, value, normativity, reasons, reactive attitudes, moral sentiments, punishment, consequentialism, Peter Strawson, Joel Feinberg

Contents

Chapter 1

Introduction

Many, perhaps most, writings on responsibility, especially on moral responsibility, approach the subject to find answers to metaphysical questions. The focus of inquiry common to these writings has been, first, the set of necessary and sufficient conditions under which one can be said to be a morally responsible agent, and second, whether human beings meet these conditions.

True, this inquiry has to a great extent been driven by the concern about responsibility. The way this concern is often put is that the ”freedom worth wanting” is the freedom without which we could not be morally responsible agents. It is less often explained, however, why we would want to be responsible agents at the first place. Why does responsibility, our own and that of others, matter to us? Why does the possibility worry us that so long as determinism is true (or false), our view of ourselves and others as responsible agents may turn out to be an illusion? What is the practical significance of ascribing responsibility? And if ascribing responsibility does have practical significance, how can we justify it?

It is perhaps safe to say that only a minority of works deal with these questions, or at least, that only a minority is more interested in them than in the metaphysical questions aforementioned. But whether or not this is a correct assessment of the literature on responsibility, it is certainly true that these are the questions this work would like to give priority to.

I believe that it is possible to focus on these questions without presupposing an answer to the metaphysical debate. Accordingly, the standing assumption throughout this work will be that the necessary and sufficient conditions under which one can be said to be a morally responsible agent can in fact be met. We have enough freedom to qualify as responsible agents– either because compatibilism is true, i.e. determinism does not undermine responsibility-entailing freedom, or because libertarianism is true, i.e. determinism is false and indeterminacy does not undermine responsibility-entailing freedom.

I will not argue in favour of this assumption. But those who are skeptical of its truth, are invited to turn everything that follows into a conditional and ask: what *could* we say about the practical significance and justification of responsibility-ascriptions if the standing assumption were true? That exercise is not entirely futile even for those who think that determinism, or the absence of it, rules out responsibility since it is worth knowing more about what sort of agency determinism, or the absence of it, makes impossible. Moreover, understanding what responsibility-entailing freedom would be like may perhaps help to understand too what it is about determinism (or the absence of it) that makes such freedom impossible thereby potentially strengthening the argument of the skeptics.

What I will argue for, however, is the claim that ascriptions of responsibility are first and foremost judgements. That priority is both normative and descriptive. Firstly and most importantly, normative because it is by virtue of being judgements that ascriptions of responsibility can have practical significance for us, i.e. they can guide our behaviour and shape our relationship to other agents. But it is also descriptive because that priority explains the psychology of responsibility-attributing practices. That is to say, in our everyday interactions too we take ascriptions of responsibility to be based on judgements and this is what renders them effective in motivating certain forms of (inter-)action. I will not deny that ascriptions of responsibility are often expressed in the form of emotional reactions, characteristic behavioural patterns and even typical sanctions. But I will try to show that these manifest responses too can only have practical significance if they are ultimately derivable from judgements.

It follows that their being judgements is the key to explain the link of responsibility-ascriptions to a whole array of important concepts with applications inside and outside morality, such as punishment, guilt, resentment, apologizing, forgiveness, restitution, praise, deterrence, and many more. I will explore some of these conceptual connections in the following chapters from this judgement-based perspective.

Taking that judgement-based perspective also entails the claim that the appropriateness of the judgement of responsibility is necessary for the appropriateness of a distinct range of manifest responses. The justifiability of these manifest responses will depend on whether or not the agent is indeed responsible. Being responsible in this sense is prior to holding responsible.

I will call the view just described the Priority Thesis. The first task, then, is to argue for the Priority Thesis and pinpoint, as far as possible, the weaknesses of rival views. The second task is to show that a theory of responsibility which adopts the Priority Thesis can indeed account for the practical significance of responsibility-ascriptions, i.e. explain and justify their normativity, their reason-giving force.

The most familiar solution to that second task is to derive the normativity of responsibility-ascriptions from the normativity of the moral requirements. I will refer to this solution as the Package Deal Argument. According to that argument, *moral* responsibility is ascribed for the violation of *moral* requirements. Therefore, the normativity of responsibility-ascriptions derives from the normativity of moral requirements. If we have reason to accept moral requirements as binding, so we have reason to judge agents in terms of how far their actions meet or violate those requirements. We are committed to the practice of responsibility-attribution because we are committed to some moral principles and their requirements.

I find that solution problematic. So I will try to point out its shortcomings and propose an alternative instead: the Value Thesis. The Value Thesis stipulates that ascriptions of responsibility track a value, the value of being a person capable of recognizing and acting on reasons. On this alternative conception, the normativity of responsibility-ascriptions is to be traced back to what we value about people. If that is true, responsibility-ascriptions generate reasons for us not only insofar as they trace the violation of moral requirements. Rather, their normativity has to do with the value of responsibility as an aspect of personhood.

This alternative conception draws support from a negative and a positive consideration. The *negative* consideration is the failure of the Package Deal Argument to satisfyingly anchor the normativity of responsibility-ascriptions in the normativity of moral requirements. It is clear why the fact that we recognize a moral principle and its requirements as valid would give us reasons to want those requirements met. But it remains unclear on the Package Deal Argument why we would be interested in the question whether those requirements are *voluntarily* met or not. But, as I will argue, ascriptions of responsibility presuppose voluntariness. This is why we are justified to respond to agents in certain ways if and only if they are responsible. However, the Package Deal Argument fails to explain why judgements specifically tracking voluntariness should have normative force. Or so I will argue.

The *positive* consideration in favour of the Value Thesis is that we ascribe responsibility not only for the violation of moral requirements, but for other things people voluntarily do as well. The example provided will be ascriptions of responsibility for the choice of one's life-plan.

But why should we value responsibility as an aspect of personhood? What justifies our commitment? I will try to answer that question by arguing that being a responsible agent is a valuable aspect of personhood because being a responsible agent is *constitutive* of what it is to be a person. It is impossible to say why that should be so without a theory of personhood itself. I will not attempt to present such a theory here. However, in closing I will offer two, mutually not incompatible, ways of showing *why* responsibility is a valuable aspect of personhood.

Let me finally outline how the following five chapters will argue for these main points. Chapter 2 proposes a definition of the concept of responsibility and seeks to defend that definition against various objections. It analyzes the concept of normative consequence and argues that a distinct range of normative consequences (e.g. punishment) requires the agent's responsibility in the sense that we have *pro tanto* reasons to impose them if and only if the agent is responsible. It also makes a case, however, for the claim that another, well-circumscribed range of normative consequences does not require the agent's responsibility. The difference has everything to do with the fact that ascriptions of responsibility presuppose that the action was voluntary. The independent normative significance of voluntariness is also defended. Through these argumentative steps, this chapter establishes the claims which jointly make up the Priority Thesis: the priority of cognitive content over emotional response and the priority of being responsible over holding responsible.

Chapters 3, 4 and 5 take a closer look at major theories of responsibility all of which pose a challenge from different angles to the judgement-based account of responsibility I would like to defend here, and specifically to the Priority Thesis. By exposing what I believe to be the shortcomings of these theories I hope to mount a negative argument in favour a judgement-based account of responsibility. This negative argument is based on the thought that for different reasons these alternative theories seem unable to do what a judgement-based account is able to do, namely to account for the normativity of responsibility-ascriptions. Chapter 6 then will put forward the Value Thesis in an attempt to show how a judgement-based account can do just that: explain and justify the normative force of responsibility-ascriptions.

In any case, beyond putting forward the criticisms summarized below, each of these chapters will also seek to capture the most important insights about responsibility (and occasionally about other related philosophical topics) which are articulated by these theories. These are important because I believe they should be taken into account by a judgement-based account of responsibility as well.

Chapter 3 focuses on the consequentialist theory of responsibility, which although judgement-based, does not observe the priority of being responsible over holding responsible. The result is that the consequentialist notion of responsibility is both implausibly anaemic and self-contradictory. This is because consequentialism not only lacks a robust understanding of ascriptions of responsibility as addressed at the agent for his action, but is also self-contradictory: the ascriptions of responsibility as consequentialists understand them are unlikely to promote the forward-looking concern which on this understanding could alone justify them.

Chapter 4 deals with an influential view of responsibility, most powerfully presented in the work of Peter Strawson, but also shared by others. I criticize the Strawsonian view for also confusing our reasons for judging that the agent is responsible with our reasons for responding to him in certain ways. In addition, I contend that the Strawsonian view also suffers from an excessive emphasis on the emotional manifestations of ascriptions of responsibility. Because it does so, it is unable to account for the normative force of responsibility-ascriptions, i.e. explain why we make such ascriptions and how we justify them.

By contrast, Chapter 5 takes on a theory of responsibility, the Ledger View, that accepts the Priority Thesis. Due to its implausible and excessively demanding conception of what justifies judgements of responsibility, however, the Ledger View entails a skeptical conclusion with regard to the applicability of the concept of moral responsibility. I will argue that that skeptical conclusion can be resisted once we realize, first, that the conception of what counts as a 'fact' adopted by the Ledger View is mistaken, and second, that the justification of judgements of responsibility depends not on so-called 'brute' facts of the physical world, but rather on facts such as the existence of practical norms or values.

In Chapter 6 then, I disambiguate the above disjunction and, by presenting the Value Thesis, make the claim that ascriptions of responsibility answer to a value (rather than to norms), namely the value of responsibility as an aspect of personhood. The problem of the normativity of responsibility-ascriptions is articulated. Then the most common way of handling it, the Package Deal Argument, is compared and contrasted with the Value Thesis. After answering some objections to the Value Thesis, I close by offering two ways to make good the claim that responsibility is an aspect of personhood to be valued.

Chapter 2

Ascribing Responsibility

This chapter is intended to establish the essential conceptual distinctions and definitions on which subsequent chapters will rest. Not all of these definitions and distinctions will be of equal importance in what is to follow. I believe, however, that it may be important to outline them here so that the position of this work can be more easily related to alternative approaches to the problem of moral responsibility.

2.1 Varieties of responsibility

The term 'responsibility', or even the more specific expression, 'moral responsibility', is used in a number of senses. Throughout this work, I will be almost exclusively concerned with only one of these, i.e. what is often referred to as retrospective moral responsibility. Other, partly related senses include:[1]

1. *Role responsibility:* It is in this sense that parents are said to be responsible for looking after their children or a captain is said to be responsible for the safety of his ship and the passengers on board. Certain functions, roles and offices impose special moral requirements on the behaviour of those who assume them. Most importantly, such people are expected to perform a more or less precisely circumscribed set of duties and comply with specific rules. These functions, roles and offices, however, need not be voluntarily acquired, e.g. brothers and sisters may be role responsible for each other merely by virtue of being siblings of each other.

   There are various complications in connection with role responsibility that I will sidestep here. Most importantly, there is the question

[1] Some of these as well as further ones, not mentioned here, are listed in Hart's taxonomy of the senses of responsibility, see Hart, 'Postscript: Responsibility and Retribution,' 210-30.

   how pressing is the duty to comply with the rules and requirements associated with a certain role. What happens when the moral requirements applying to the agent by virtue of his assuming a certain role clash with other moral or non-moral requirements also shouldered by that agent? Where exactly are we to draw the line between a simple duty arising from the circumstances (e.g. the duty of easy rescue) and role responsibility which usually, but certainly not always, involves a number of complex duties over a longer period of time?

   In any case, the crucial point is that role responsibility is generally taken to stake out a set of special reasons for action that the agent whether he assumes the role voluntarily or involuntarily: The captain of the ship *qua* captain has strong reasons to act in the interest of his passengers' safety, parents *qua* parents have strong reasons to act for the benefit of their children, and so on. It is also clear that this usage is not unrelated to the sense of moral responsibility I will be mainly concerned with in the following. This is because having role responsibility implies that the agent can be called to task for failing to act in accordance with certain moral requirements associated with the function, role or office in question.

2. *Responsibilities:* The word responsibility is frequently used synonymously with duties and obligations. One often comes across statements such as: 'It is the students' responsibility to return books to the library in time', or 'citizens not only enjoy rights but also shoulder responsibilities towards their government'. In this sense, to have a responsibility means simply to incur a duty or obligation. But this terminology can be confusing. Strictly speaking, what an agent is held responsible for is having discharged or having failed to discharge a duty or obligation (and possibly many other things). Adhering to this distinction, I will not to use the word responsibility to mean duty or obligation in the following.

3. *Capacity-responsibility:* This sense of the term 'responsibility' focuses on the criteria of responsible agency. Every agent has to meet certain conditions to count as fit to be held responsible. Ignorance, coercion, duress, mental illness, infancy, etc. are usually regarded as exculpating factors (a lot more about these factors will be said later on). In any case, if the agent is said to be responsible in the capacity sense of responsibility, then it is understood that none of these responsibility-undermining conditions obtain.

Note that the last understanding of the term appears to correspond most closely to the etymological origins of the word 'responsibility' (derived from the Latin equivalent of the verb 'respond', i.e. 'answer'). To regard people as responsible in this sense is to see them as capable of 'answering for' their actions. The possession of these capacities is important because it is these capacities that enable the agent to deliberate and to act in accordance with (or act for) reasons.

It seems less persuasive, however, to say that when an agent is ascribed (moral) responsibility for her actions, then *all* that is involved in such an ascription is that she possesses these capacities. What I want to argue in the next section is that while having these capacities may indeed be a necessary condition of ascriptions of moral responsibility, it does not exhaust the *content* of such ascriptions.

2.2 The Ascription Thesis

How are we to characterize ascriptions of responsibility? In a sense, answering this question is all that the following work endeavours to do. Clearly, there is considerable disagreement with regard not only as to what the right answer may be but even as to what may qualify at all as an answer.

I will propose to understand an ascription of responsibility as a normative judgement. That is to say, when we ascribe responsibility we judge the nature of the action and that judgement generates specific reasons for action. I will argue that ascriptions of responsibility can provide the reasons for a distinct class of actions. I will refer to actions of this class as the normative consequences of responsibility-ascription. My claim is that we have reasons to perform actions of this class because and only because an agent is judged to be responsible for his action. I will also argue, however, that ascribing actions to agents without necessarily ascribing responsibility to them for these actions can also have normative consequences. These are the normative consequences of action-ascription.

Ascriptions of responsibility are defeasible. This means that a judgement of responsibility also entails the belief that no responsibility-undermining conditions obtain. Responsibility-undermining conditions can be global (exemptions) or local (excuses). If a global responsibility-undermining condition obtains (e.g. the agent suffers from severe mental illness), then the agent is not capacity-responsible. If a local responsibility-undermining condition obtains (e.g. coercion, non-culpable ignorance), then the agent is not responsible for that particular action.[2]

An ascription of responsibility is normative in the sense that it has reason-giving force. The reasons thus given need not be exclusively *agent-focused,* i.e. reasons for actions to be done *by* or to be done *to* the person to whom responsibility is ascribed. Suppose I judge you responsible for Φ-ing.

[2] In labeling global responsibility-undermining conditions as 'exemptions' and local ones as 'excuses', I follow Gary Watson. See his 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,' 259-61.

That judgement gives you reasons for various kinds of action towards me and others, gives me reasons for action towards you, but it may give me reasons for action independently of you as well (e.g. I may just be justified in venting my resentment even if you will never hear it). And further, it may give someone else reasons for action too, someone who was not affected by the actual consequences of the action at all (apart from learning about it).[3] Nevertheless, the normative consequences, for which ascriptions of responsibility give us reasons, are frequently imposed on the agent himself.

It follows from the above that to say that to judge an agent responsible is to judge that the agent is liable to special normative consequences because of his action. But since ascriptions of responsibility are defeasible, so is the notion of the agent's liability. Therefore:

> *An agent is responsible, if he is subject to normative consequences for* Φ*-ing and no responsibility-undermining conditions apply.*[4]

But note that an ascription of responsibility can be made in response to a justifiable or unjustifiable action. Therefore:

> *If* Φ*-ing is not justifiable all-things-considered (e.g. because it is morally wrong), then* X *is blameworthy for* Φ*-ing. If* Φ*-ing is justifiable all-things-considered (e.g. because morally required), then* X *is praiseworthy for* Φ*-ing.*

I will refer to these set of claims defining ascriptions of responsibility as the Ascription Thesis for short.

The first point to make about the Ascription Thesis is that the term 'normative consequences' is a general term intended to cover everything that an ascription of responsibility to an agent gives anyone (including the agent himself) *pro tanto* reason to do or to feel. Thus an ascription of responsibility can furnish us with *pro tanto* reasons for such diverse responses as overt or covert moral criticism, punishment, guilt and resentment, the making of an apology, and so on. Many normative consequences are duties which the agent incurs as a result of his Φ-ing such as the duty to make compensation or to apologize. On the whole, I will try to show that many theories of moral responsibility take a too narrow view as to what can constitute a normative consequence of an ascription of moral responsibility.

But why *pro tanto* reasons? Because an ascription of responsibility is never a sufficient condition for the imposition of normative consequences.

[3] For the purposes of this discussion, reasons for action are also taken to include reasons to feel something such as guilt or resentment.

[4] The variable Φ can also stand for the agent's *being* in a certain way. I will try to show this in Section 2.5. I argue here that the imposition of normative consequences can also be justified as a response to the agent's character.

Thus 'being responsible' is to be distinguished from 'holding responsible'. This is perhaps the most basic distinction to be made for the purposes of this work. On the one hand, as already indicated, I will argue that an ascription of responsibility is first and foremost a judgement that the agent *is* responsible for something he has done.[5] As proposed in the definition above, this judgement entails specifically the beliefs that:

1. *no global responsibility-undermining conditions obtain, i.e. the agent is capacity-responsible;*

2. *the action is justifiable (if praiseworthy) or not justifiable (if blameworthy);*

3. *no local responsibility-undermining conditions obtain, i.e. no excuses (e.g. compulsion or ignorance) apply to the particular action which could absolve the (otherwise capacity-responsible) agent from responsibility.*

However, no judgement of responsibility, even if entirely justified, entails on its own that any given overt response is justified.[6] For example, a judgement of responsibility may justify the imposition of punishment on the agent or voicing one's resentment to him, but other conditions must also be satisfied for it to be right to do so (e.g. one must be authorized or entitled to impose punishment).

On the other hand, it is true that ascriptions of responsibility are necessary for a distinct range of normative consequences. Thus I disagree with the view that the imposition of *all* normative consequences can be justified independently from the appropriateness of the judgement concerning the agent's responsibility. So I will also argue that specific normative consequences are predicated upon the correctness of the judgement that the agent is worthy of blame or praise.

More controversially perhaps, I also wish to say that an agent can incur certain normative consequences without being responsible for those actions. In general, being worthy of blame for the action is not necessary for the imposition of some negative normative consequences. For example, the agent may justifiably incur the duty to make restitution for an action which itself was justified or excused. By the same token, being worthy of praise is not necessary for the imposition of some positive normative consequences.[7]

[5] Or the way he is, see Section 2.5.

[6] I will argue this point in more detail in Chapter 3 as the principal criticism against consequentialist theories of moral responsibility.

[7] Unless otherwise indicated, I will henceforth treat worthiness of praise and worthiness of blame, the justification of positive and negative normative consequences, etc. as symmetrical. Not everyone would agree with this, see for example Smart, 'Freewill, Praise, and Blame,' 303-5.

Note, once again, that being judged responsible for a particular action is not to be equated with capacity-responsibility. Having the requisite capacities is necessary for the agent to be responsible, but over and above the identification of certain capacities, the agent's responsibility also entails that she is liable to incur various normative consequences for her Φ-ing.

What the Ascription Thesis says is that ascriptions of responsibility involve a certain kind of appraisal, the establishing of a special kind of link between the agent and the rightness or wrongness of the action and this is where the practical importance of these ascriptions derives from. Of course, the obtaining of certain conditions can block the route leading from right or wrong actions to ascriptions of responsibility. Thus, as already mentioned, the possession of certain capacities by the agent is required for the ascription of responsibility to be appropriate. If there is reason to believe that the agent does not possess these capacities in the requisite degree (due to some serious mental disorder, say), then no responsibility will be ascribed to the agent for whatever he may do. Similarly, if the agent was relevantly hindered in exercising these capacities (for instance, because he was exposed to physical coercion), then the agent is excused from bearing responsibility for that particular action.

That certain conditions relating to the agent's capacity or the circumstances of the particular action can defeat the ascription of responsibility is indicative of why responsibility-ascriptions have normative force. Specifically, focusing on the conditions which are commonly taken to block ascriptions of responsibility raises two questions.

First, there is the question what requisite properties or necessary circumstances for responsibility to be ascribable are missing when a responsibility-undermining condition obtains. For the most part this work concentrates on the normative implications of being responsible rather than on the metaphysical pre-conditions of being a responsible agent. Therefore, I will not be able to give detailed attention to the first question in this work.[8] This question is inseparable from metaphysical issues concerning the pre-conditions of responsibility-entailing freedom. In particular, it is inseparable from the thesis of determinism. This is also why I cannot discuss the question whether all excuses and exemptions are taken to undermine responsibility for the same reason, e.g. the reason that they all indicate that the agent could not have done otherwise. I cannot enter into this question because to do that I would have to take a stand on the metaphysical issue whether or not determinism deprives the agent of alternative courses of action.[9]

[8] Except for a digression in Section 4.5. The connection between excuses/exemptions and determinism is also touched upon in Section 3.2.

[9] Cf. Wallace, *Responsibility and the Moral Sentiments,* esp. Chapters 5 and 6, for an alternative account of excuses and exemptions. Wallace does not rely on the assumption that excuses and exemptions are recognized because they indicate the lack of alternative possibilities available to the agent, but proposes that we justify excuses and exemptions on

Second, we also need to ask why the presence of certain feature(s) or circumstance(s) are understood to be necessary for a responsibility-ascription to furnish us with valid reasons for action. Why do we think that certain things must be true of the agent and the circumstances of the particular action for the agent to bear responsibility for that action and be liable to normative consequences because of it? For example, why do we think that certain things must be true of how the action was carried out by the agent, for us to have *pro tanto* reasons to punish the agent for it? I will discuss this question in more detail below as well as in Chapter 6.

Finally, two important points need to be added here about the notion of responsibility supported by the Ascription Thesis. It follows from the Ascription Thesis that ascriptions of responsibility are *backward-looking* and that they are *individualized.*[10]

Thus the Ascription Thesis stipulates that responsibility is ascribed for doing wrong or right in the past.[11] This is also in line with how ascriptions of responsibility are commonly understood in everyday practice. But the significance of this definition will emerge fully only in the discussion of consequentialist theories of responsibility. To anticipate, the crucial point is that since ascriptions of responsibility are focused on the agent's having done wrong (or having acted rightly), the purpose of ascribing actions to agents cannot simply consist in determining the likelihood of the agent acting in similar ways in the future. We are, for reasons to be discussed in subsequent chapters (and especially in Chapters 3 and 6), concerned with what the agent has done irrespective of whether or not the character of the action bodes ill or well for the future. And if that is true, then it is also right to say that we do not impose normative consequences on the basis of that judgement because we are motivated by a forward-looking concern (to maximize utility or to minimize the violation of rights). In sum, insisting on the backward-looking nature of responsibility-ascriptions is to insist that the

"quite different principles of fairness", ibid., 116. These principles he argues are compatible with the truth of determinism.

[10] Wallace, *Responsibility and the Moral Sentiments,* 56.

[11] Or possibly in the present if the wrongdoing is simultaneous with the ascription. Can we ascribe responsibility for what the agent is going to do in the future rather than what she has done in the past? In a sense, we can. I can blame you today for not turning out to the meeting tomorrow. But that is either because you give every sign of having formed the intention not to go or because it is already obvious that you will not be able to make it to the meeting although she should be there. In both cases, however, responsibility is obviously ascribed not for the future action but for something you do or fail to do in the present or past: your forming a reprehensible intention or her culpable negligence (assuming of course that this is your fault and it is not due to a circumstance beyond your influence that you will not be able to go). In that sense, ascriptions of responsibility can be seen as responding to what people have done or intend to do.

normativity of these ascriptions, their being a source of *pro tanto* reasons for action, is not dependent on our interest in bringing something about.[12]

By the same token, ascriptions of responsibility are *individualized.* This means that the point of blame or praise-entailing ascriptions is not merely to place wrong or right actions on somebody's doorstep in order to determine who should bear the consequences of wrongdoing, the losses and harm suffered, or who, if any, should benefit from rightful actions. Ascriptions of moral responsibility to link actions to the particular agent who has done wrong or right, and to no one else. This may sound trivial but, as will be seen, it is questionable whether consequentialist theories can explain why being responsible (as opposed to bearing the costs of action) is non-transferable.

2.3 Two objections: the 'capacity view' and the 'control view'

At least two important objections can be raised against the account of responsibility ascriptions I advocate here. The first objection is based on what I will call the *capacity view of responsibility.* The capacity view is fully satisfied with seeing ascriptions of moral responsibility as establishing the possession of certain capacities in the agent–what I took above to be only one of the necessary conditions for responsibility to be ascribable. What proponents of this view find disturbing about the Ascription Thesis is that it supposedly equivocates between being responsible and the various normative consequences of ascriptions of responsibility. The worry is that if responsibility is defined as entailing *pro tanto* reasons for action, then one often does best to deny one's responsibility claiming that an excusing condition obtained impairing one's capacities requisite to being a responsible agent: "[. . .]one's responsibility is something to be regretted and (if possible) avoided. For it is none other than a vulnerability to adverse normative consequences in the event of wrongdoing".[13]

But far from being something regrettable, it is protested, to be a responsible agent gives one the right to be recognized as a fit subject of appraisal at the first place. The Ascription Thesis is unconvincing, it is said, because the purpose of ascriptions of responsibility, far from serving to mete out unwelcome normative consequences, is to acknowledge the agent's status as a member of the community of rational beings, as someone with whom we can reason together. It follows that responsibility is something to be proud of,

[12] See Korsgaard, 'The Reasons We Can Share: An Attack on the Distinction between Agent-Relative and Agent-Neutral Values,' 275: "I believe that. . . a basic feature of the consequentialist outlook still pervades and distorts our thinking: the view that the business of morality is to *bring something about."*

[13] Gardner, 'In Defence of Defences,' 259.

even if the normative consequences of responsibility ascriptions are indeed quite often unpleasant for the agent.

This is shown, it is added, by the fact that even when an excusing condition is actually cited to forestall the imposition of adverse normative consequences, the excuse will often highlight that the agent had reasons to entertain certain beliefs or emotions, even if those beliefs and emotions themselves turn out to be ultimately unfounded or false: "The defendant did not have reason to kill her husband, for instance, but she certainly had reason to be so terrified by his obnoxious behaviour that night that she was driven to kill him."[14]

Now, it has already been accepted that the possession of certain capacities is presupposed by responsibility-ascriptions. If the agent did wrong (or right) and he is demonstrably in possession of the requisite capacities then, and only then, is it appropriate to attribute responsibility to him for that action. The capacity view rightly emphasizes that responsibility-ascriptions, by being predicated on the possession of these capacities, involve the recognition of the agent's status as a rational deliberating subject. Indeed acknowledging the agent's status forms an important part of the kind of appraisal that responsibility ascriptions represent. The agent is blamed or praised precisely because it is assumed that he is (or would have been) capable of acting on reasons. That is the important truth contained in the capacity view of moral responsibility.[15]

This truth, however, is easily accommodated by the Ascription Thesis. At the same time, the capacity view makes a number of questionable assumptions. For one thing, responsibility ascriptions are, as the above definition makes clear, dichotomous: if the agent is responsible, then she is *either* blameworthy *or* praiseworthy. So, quite simply, it is not clear why the objection concentrates exclusively on the *adverse* normative consequences of responsibility ascriptions. Attributions of praise can without doubt have positive normative consequences of great material and psychological significance as well. But if that is correct, then the Ascription Thesis does in no way portray moral responsibility as "something to be regretted and (if possible) avoided".

Second, there is a deeper worry about the alternative view of responsibility defended by those who embrace the capacity view. The problem is that on the capacity view the normative force of responsibility ascriptions for particular actions threatens to pale into insignificance. This fails to do justice to the central evaluative and action-guiding function of responsibility-ascriptions. When ascribing responsibility to an agent, one does more than just establish that that agent possesses certain capacities however important

[14] Ibid., 260.

[15] I will say more about the role of responsibility-ascriptions in constituting the status of personhood in Section 6.3.

these capacities may be. Ascriptions of responsibility, by linking actions to agents, constitute a normative judgement. They establish not only that the agent had the requisite capacities to do this or that but also, to anticipate a more detailed discussion to follow later on, that the agent was the kind of person who did this or that.

So contrary to the capacity view, I believe we must say that the meting out of a distinct range of normative consequences cannot be divorced from the agent's responsibility. Certain responses to agents have a special force not because they entail harsh or unpleasant treatment or any concrete setback of the agent's interests, but *only because* they single out the agent as responsible for what he has done (such is the case, for instance, with expressions of reactive sentiments such as resentment or indignation).[16] Indeed, I will try to show that ascriptions of moral responsibility are required for the justifiability of imposing certain normative consequences. These normative consequences differ from those imposed on the agent for actions without a corresponding judgement of responsibility.

Thus perhaps the most important normative consequence restricted to blameworthy actions is punishment.[17] There is a good case to be made that blameworthiness is a necessary (though not a sufficient) condition of liability to punishment. It is said that its connection to blameworthiness rather than a difference in the degree of severity is what really distinguishes punishment from mere penalties.[18] Note, however, also that in addition to requiring blameworthiness punishment also seems to involve the idea that it is not unjustified to subject the offender to 'harsh treatment', whereas other normative consequences also presupposing blameworthiness do not involve such treatment. The fact that there is a considerable variety as to what justifies the imposition of these normative consequences in addition the agent's responsibility seems to support the claim made earlier that responsibility is a necessary but not a sufficient condition of holding the agent responsible, i.e. of the actual imposition of normative consequences.

The *control view* attacks the Ascription Thesis from a different direction. It holds that that an action cannot be judged right or wrong if no agent can be assigned responsibility for it: "In brief, the rightness of actions depends on the possibility of holding people responsible and crediting them with responsibility".[19] Clearly, if this was true the Ascription Thesis would be rendered circular. If the assessment of doing wrong or right depended on ascriptions of moral responsibility, then defining moral responsibility in

[16] More on reactive sentiments in Chapter 4.

[17] It is worth noting that representatives of the capacity view also admit that punishment cannot be imposed unless the agent is blameworthy, see Gardner, 'In Defence of Defences,' 259 and Gardner and Macklem, 'Reasons,' 468.

[18] Feinberg, *Doing and Deserving,* esp. 98-101.

[19] Honderich, *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes,* 60.

terms of the agent's blameworthiness or praiseworthiness for doing wrong or right would get us nowhere.

I would like here to steer clear of the difficult question of how exactly to characterize the rightness and wrongness of actions. But it appears that on most definitions there is no reason why we should accept the claim on which the second objection is based. Wrongdoing, for instance, may be defined simply as causing harm or loss to someone. Frequently, however, such harms and losses are not intentionally caused, for instance when they occur as the unforeseeable consequences of certain actions. What's more, they may even be intentionally caused without thereby making the causing agent morally responsible. We are dealing with such a case for instance when the harm or loss is inflicted as the unavoidable negative side-effect of an action carried out under necessity, duress, etc. Or possibly, wrongdoing may be defined as the violation of a person's claim-rights. But again these rights violations may not be deliberate or foreseeable, or they may even be deliberate and yet not make the causing agent responsible for them.

It will be found that the control view is based on a more comprehensive approach to morality. The following citation sums up this view in a succinct fashion: "Morality is concerned with the world in so far as it is in our control. An action for which the agent cannot be assigned responsibility, and which he therefore cannot be taken to control, is an event which falls outside of morality's concern or province".[20] In fact, this bald assertion seems to be the only point invoked in support of the objection that actions cannot be right or wrong without someone incurring responsibility for them. Now, it could at this point even turn out to be right, and some authors do in fact hold the view, that one cannot speak of wrongdoing without blame-entailing moral responsibility (although I, for one, certainly do not think so and will argue against this in subsequent sections). In any case, however, the correctness of that view certainly cannot be shown by simply assuming that "morality's province" does not extend beyond actions in our control. There is nothing immediately compelling about this view of morality, nor anything counterintuitive in denying it.

2.4 Ascriptive theory

According to one possible classification, practical philosophy divides into three domains: value theory, normative theory and ascriptive theory.[21] Value theory studies fundamental values such as what is good and what is bad, normative theory deals with the rightness and wrongness of actions,

[20] Ibid., 59.

[21] Such a three-fold distinction is made in Raz, *Practical Reason and Norms,* 11-12 and a similar one in von Wright, *Norm and Action,* 6-7. Note that this tripartite distinction can be applied not only in ethics, but in other areas of practical philosophy as well such as aesthetics or law for example.

and ascriptive theory focuses on the preconditions for and the implications of attributing actions to agents. There are several ways of establishing connections among these three domains. For instance, on consequentialist accounts normative theory is taken to presuppose a theory of value insofar as those actions will be deemed right which maximize a certain value. Rawls, on the other hand, argued that normative theory should be seen as prior to the theory of value. On his theory, what is good is to be defined in terms of what is right.[22]

It is sometimes argued that ascriptive theory is independent from value theory and normative theory, at least as far as ascriptions of responsibility are concerned. We have various criteria for evaluating what actions are right and what actions are wrong. With this account of wrongdoing/rightdoing in hand, we can then proceed to investigate the conditions under which it seems justified to attribute right or wrong actions to agents. Doing so is important because our attribution of a particular action to a particular agent may change the way we relate to or treat that agent. But, and this is the salient point, identifying the criteria which have to be met by the agent if we are to attribute the action to him is not dependent on our commitments to certain values or norms.[23] I will call theories which accept this argument 'objectivist' because what they hold is that as long as the agent instantiates certain objective properties, it will be correct to attribute the given action to him and impose the concomitant normative consequences too.[24]

My proposed alternative to objectivist theories will be spelled out in Chapter 6. But let me explain here why I think objectivism is unattractive. It seems to me that ascriptions of responsibility, far from being independent from them, are actually governed by the norms or values we accept.

The first thing to note is that the objectivist account leaves unexplained why we should think of excuses and exemptions as undermining the agent's responsibility. This is a problem because nothing in the familiar and broadly acknowledged responsibility-undermining conditions themselves–e.g. ignorance, duress, mental disturbance–determines how are we to respond to actions performed under those conditions. If the execution of the mentally ill by our Victorian predecessors or American contemporaries is wrong, it is not because it involves a factual or logical mistake. We think that doing so is somehow wrong or unfair.

The objectivist objection to this is that a "normatively detached outsider can still tell whether someone is fit to be held responsible" and therefore

[22] Rawls, *A Theory of Justice,* 24.

[23] I have here in particular Philip Pettit's theory of freedom and responsibility in mind. See Pettit, *A Theory of Freedom: From the Psychology to the Politics of Agency,* esp. 26-7.

[24] See for instance ibid., 4: "To be free, in the most general sense, is to be fully fit to be held responsible; it is to be fully deserving of the sort of reactions, say those involving resentment or gratitude, that characterize face-to-face relations".

specifying the conditions of attributability cannot be dependent on our commitment to norms or values.[25] While it is certainly true that a normatively detached outsider will be able to tell whether the norm has been applied correctly that observation seems to miss the point. The real issues are, first, what facts about the agent and the circumstances of the action should be taken as pertinent to our judgements of responsibility, and second, what those judgements give us reason to do. Both of these issues are inseparable from our acceptance of certain values or norms.

Thus note that there can be disagreements concerning the scope of pertinent facts as well. Someone may think that addiction is not an excusing factor because addicts do not really lose control over their action to a degree that would warrant excusing them from responsibility for those actions. Others criticize the widely-held view that severe childhood deprivation or abuse is generally to be accepted as excusing from responsibility.[26] These matters may be discoverable by means of empirical research. The crucial point is, however, that even if all the facts are in, the justification of ascribing responsibility on the basis of those facts will still be dependent on what we believe to be right/wrong or what we value.

So I think that the objectivist account of responsibility is implausible. A discussion of the way in which I believe responsibility-ascriptions to depend on our commitments to value will be postponed until Chapter 6. It is also there that I will attempt to explain why I think responsibility-ascriptions depend directly on our commitment to the value of responsibility rather than on our commitments to moral norms.

2.5 Varieties of normative ascriptions

Some argue that we cannot be liable to normative consequences for our actions unless those actions are voluntary. The position I would like to defend here is that we cannot be responsible for our actions unless they are voluntary, but that we can be liable to normative consequences even for what we non-voluntarily bring about or for the way we non-voluntarily happen to be.

Needless to say, there is considerable disagreement in the literature as to how to spell out what voluntariness consists in. I do not presume to be

[25] Ibid., 27.

[26] See Moody-Adams, 'On the Old Saw that Character is Destiny.' Moody-Adams disputes this common defence because she rejects as psychologically unrealistic the claim that victims of childhood deprivation experience truly irresistible impulses to aggressive behaviour, cf. esp. 123-4. Her point may or may not be right (the majority of authors on this subject seem to agree that she is not right). But note that even if we knew for certain that she was wrong, i.e. that childhood deprivation does in fact lead to irresistible impulses, it would still be an open question whether this fact should be taken as relevant to our ascriptions of responsibility or not.

able to answer this question. Especially, I will not venture an answer as to whether voluntariness in the requisite sense is or is not compatible with the truth of determinism. But, and this is all we need to establish at this point, it seems that most compatibilists and incompatibilists agree that for the agent to be responsible, and *a fortiori* to be liable to certain kinds of normative consequences predicated on the agent's responsibility, some kind of voluntariness is required.[27]

I will not question this tenet. What I want to take issue with here is the view that what may be legitimately done to us or what may be thought of us in consequence of what we have done must fully coincide with the extent of our voluntary control over our actions and their consequences. Two questions arise at this point: *First,* why is voluntariness required for ascriptions of praise-entailing or blame-entailing responsibility to be appropriate? *Second,* is it true that voluntariness is not a pre-condition for the imposition of various, possibly quite adverse, normative consequences on agents in consequence of what they have done?

2.5.1 Two type of appraisals: 'ought-to-be' and 'ought-to-do'

Let me begin with the second question. Consider the following statements:

1. John ought not to be so cruel.
2. Claire ought to be ashamed of herself.
3. Mark ought to be proud of his achievements.
4. Rachel ought to have remembered her father's birthday.
5. Des ought to know that torturing animals is wrong.
6. No child ought to starve.

(1)-(6) all represent perfectly standard usages of 'ought'. At the same time, (1)-(6) all refer to cases in which it is not obviously within the agent's power to change the state of affairs objected to in the individual sentences. (1) expresses moral criticism of John's character, (2)-(3) complain about the lack of certain emotions in the agent, (4)-(5) point out reprehensible cognitive failures in the agent, (6) voices indignation about a given state of affairs in the world as it is. Character, emotions, cognitive performances (or their lack) and certain states of affairs are the most prominent instances in which the use of 'ought' appears appropriate despite the fact that it is not

[27] A notable exception is Adams, 'Involuntary Sins.' See below for a response to Adams's account.

necessarily up to a particular agent (or anyone at all) to do anything about what is being objected to.

Granted, one's character or emotions or cognitive performances and even certain general states of affairs can to some extent be influenced by agents' voluntary interventions. Emotions and desires can be intensified or suppressed. Typical failures of our cognitive functioning, say notorious forgetfulness, can be avoided by adopting familiar routines and attitudes.[28] It should also be pointed out that references to one's character are often ambiguous. What is meant can be either a dominant streak in one's personality (e.g. cruelty) or our entire way of being, the sum total of all our emotional, intellectual, psychological and other characteristics. It is probably unrealistic to expect one to be able to change one's character in the latter sense, but it is not so unfeasible to believe that one is capable of reforming one's character in the former. And it is character in this sense that usually constitutes the object of criticism (consider again cruelty or cowardice or dishonesty). Finally, while nobody alone can put an end to child famine, an utterance such as (6) can be taken to mean that everyone ought to do something about it.

Having said that, the scope of voluntary interventions is limited in such cases. Moreover, and this is the crucial point, the 'ought' occurring in (1)-(6) will be used even if it is clearly not up to the agent to do something about what is being objected to. Thus whether or not Jones is in fact able to do away with the cruel streak in his character, we will find his cruelty reprehensible. I believe that similar considerations can be applied to (2)-(6). This has led many to distinguish between two senses of ought, 'ought-to-do' and 'ought-to-be' where it is understood that 'ought' figures in (1)-(6) in the latter sense. It ought to be the case that no child had to starve, it ought to be the case that Jones was not a cruel person, it ought to be the case that Claire felt ashamed herself, and so on. Lamentably, however, all this is not the case, nor can this be helped for the time being. By contrast, the first sense of 'ought', ought-to-do, occurs in propositions such as 'Jones ought to keep his promise' or 'Smith ought to have rescued the drowning child'.[29] It

[28] (5) raises the problem of culpable ignorance. It is assumed in this example that Des is not to be blamed for not knowing that torturing animals is wrong, i.e. he is not culpable for his ignorance. For instance, Des may have been raised in a community which does not regard animals as capable of suffering. Of course it is often the case that we do not know that *p*, although we ought to know that *p* and we could have known that *p* (had we paid attention, say). In that case we are blameworthy for not knowing that *p*.

[29] The terms 'ought-to-be' and 'ought-to-do' stand for two senses of 'ought'. These two senses are used in different kinds of appraisals as I will try to show below. Meanwhile, I would like to note an ambiguity of this terminology, taken from Zimmerman, *An Essay on Moral Responsibility*, 117 and elsewhere. Zimmerman is well aware of this, see for example Zimmerman, *The Concept of Moral Obligation*, 2-3. Here he points out that one could also say the following: 'it ought to be the case that Jones keeps his promise' or even 'it ought to be the case that Smith rescued the drowning child'. The main point is, however, clear. For the kind of appraisal that yields 'ought-to-be' statements, such as

is assumed that Jones and Smith are in a position to act voluntarily and what is expressed by the 'ought' used in this sense is that they are required to do so.

In sum, the distinctive mark of 'ought' in the sense of 'ought-to-be' as opposed to 'ought-to-do' is that the former does not necessarily presuppose voluntariness (why the latter does presuppose voluntariness will be discussed below). 'Ought-to-be' appraisals are made without regard to the question whether anyone can do anything about the situation that is the cause of irritation or contentment. For this reason it is also true that while 'ought-to-do' can be used in a morally binding sense to denote an obligation or duty ('you ought to keep promises', etc.), 'ought-to-be' does not.[30]

This distinction would probably be accepted by almost everyone *without* the further claim that 'ought-to-be' type evaluations can also constitute reasons for the imposition of normative consequences. I would like to argue, however, in favour of this additional claim too.

Some have tried to establish the normative relevance of 'ought-to-be' appraisals by showing that their normative force is in fact based on 'ought-to-do' type judgements. The idea is that utterances of the kind 'it ought to be the case that' can only make sense given the further assumption that someone, at least in principle, could do or could have done something about what ought to be the case. But I think there is no reason why we should draw the connection in this way. Even if there was nothing anybody could ever do about Jones's cruelty, neither himself nor his parents and so on, his cruel character continues to be judged negatively. In fact, even if determinism turns out to be true, 'ought-to-be' type appraisals remain as forceful as ever. The world would still be a better place without a cruel Jones and Jones would still be a better person without his cruelty.

The following objection could be made here: Neither character as such, nor a character trait, however dominant, can by itself constitute the subject of negative valuations since these things do not cause harm or make anything better in themselves. Rather, it is a cruel or coward *action* that causes harm and a generous or brave *action* that does good and hence only these actions should be evaluated (morally or otherwise) but not the character traits they

(1)-(6), it does not make sense to say that the subject of appraisal ought to do something about what is being objected to because, by assumption, she cannot help what is the case (while of course she may be able to do other things, e.g. conceal her lack of shame, make up for having forgotten the birthday, etc.). By contrast, while it may be true that *now* Smith can do nothing about not having rescued the child, he ought to have done (and by implication) could have done something about it *then.*

[30] I have spoken of the difference between the two senses of 'ought' in terms of the requirement of voluntariness. A different way of putting the same condition is to say that 'ought implies can' is true for obligation-generating 'ought-to-do', while it is not true for 'ought-to-be'. This is, for instance, how Michael Zimmerman draws the distinction between the two senses of 'ought' which he does not believe to be reducible to one another. See ibid., 3.

stem from. I accept that character is in effect a disposition to act in certain ways. At the same time, I do not accept that character understood as such a disposition cannot be evaluated in a way that would have normative consequences.

Imagine that Gutless Gilbert is a person who is known to have acted in a cowardly fashion on repeated occasions in the past. Now imagine that we have to pick a person for a dangerous but important mission. The candidates have not been consulted before but only a small number of people are eligible since special skills are required. Gutless Gilbert, although he possesses these skills, will not be considered to be a suitable choice because he is known to be a coward. The reason for the rejection lies in the negative evaluation of his character. He ought to be different, we seem to be saying, and not, he ought to have acted differently (we may think that also, but that is not why we reject him here and now). Consider this analogy: When choosing the right equipment for a hiking-trip we will not take a thermos flask with us that we know to be fragile. It may even be that flask has never broken before, but it would still be unwise on our part to take it with us if we judge that it is likely to break.

So I believe that if 'ought-to-be' type appraisals have normative import, and I think the Gutless Gilbert example shows that they do, this normative import is not derived from their being ultimately reducible to 'ought-to-do' type evaluations. But how can 'ought-to-be' type appraisals be normatively relevant and why can they even justify the imposition of adverse normative consequences *despite the fact* that they do not presuppose voluntariness? This is what is at stake here.

While many would presumably go along with the weaker claim that the two sense of 'ought', 'ought-to-do' and 'ought-to-be', correspond to two different kinds of value, some would perhaps shy away from this stronger claim. They would say that 'ought-to-be' type appraisals are alright as long as they merely establish a person's moral failing, virtuousness, etc. (or the moral reprehensibility/admirability of a certain state of affairs), but given that for 'ought-to-be' ought does not imply can, 'ought-to-be' type appraisals cannot vindicate the imposition of adverse or positive normative consequences on the agent.

The idea underlying this objection is not without intuitive appeal. After all, why should you suffer a loss of any kind for something–your congenital cowardice, your biographically rooted emotional insensitivity, etc.–that you could not do anything about? And for that matter, why should you benefit from doing something–acting out your inborn meekness, your rock solid work ethics drilled into you by your parents, etc.–that you could equally not do anything about? Surely, it would be particularly unfair to make the agent 'pay' for her failure to meet certain norms if she could truly not help doing so.

I believe that the right answer to this worry depends on what is understood by 'payment'. For instance, if the payment to be levied constitutes what is really a form of punishment, then it would indeed be unfair to subject the agent to it because it seems very plausible that punishment can only be appropriate if the punished act was voluntary.[31] But the imposition of other normative consequences on the basis of 'ought-to-be' type appraisals is not, in my view, equally objectionable. This, I think, is already clear from Gutless Gilbert's story above. Gilbert cannot object to not having been selected for the mission on the grounds of his failing. Moreover, if the selection was public Gilbert will even have to accept the social stigma of being a cowardly person. This may just be one of the adverse normative consequences of the 'ought-to-be' type appraisal. Nor can Gilbert protest that it was unfair to impose this or other adverse normative consequences on him for his unfortunate character trait since his cowardice is congenital, i.e. there is indeed nothing he can or could ever do about it (even if this is, let us assume for the sake of the argument, demonstrably true). If so, it would indeed be unfair to blame or punish Gilbert for being a coward, but it is not unfair to state that it would be better if this were not the case and it is not unfair to treat him as a coward person is usually treated, e.g. reject him as a candidate for the mission or decline his friendship.

The point becomes even clearer if the agent is judged according to norms she herself has consented to. Most competitions appear to be of this type. It is the participants' free decision to enter the race. Those who fail to gain the prize can hardly complain after the race of having been treated unfairly even if it is true that their handicap against the winner was not something they could ever do anything about. It may well be not their fault that they cannot run faster, in fact they may even have realized their own potential to a higher degree than the winner has, but unfortunately the race was about reaching the finishing line within the shortest possible time. Analogous considerations are applicable, I submit, even to situations in which moral qualities constitute the object of appraisal. It does happen fairly often that we submit ourselves willingly to evaluations of our moral qualities by others. This is what happens when one person courts another's friendship, one hopes to join a closely-knit team, or excel through one's courage in the army, one's chastity in a monastery or through one's dedication in a job. In certain cases, one also submits oneself to such a 'competition of moral qualities' when one applies for a public office. Should we fail such a test, as we often do, we cannot complain that our moral weakness was not something we could ever do anything about. It ought to be the case that our character was different, or that we were capable of feeling different emotions but unfortunately it is not so. Of course, given that ex hypothesi our moral failing is not our fault

[31] See page 20 above.

it would be right to complain about being blamed for it, but certainly wrong to complain about being subjected to 'ought-to-be' type appraisals.[32]

In sum, 'ought-to-be' type appraisals stand for an independent kind of normative evaluation, not reducible to 'ought-to-do' type judgements. That 'ought' can be normatively relevant even if it is used in the sense of 'ought-to-be' is shown by the fact that 'ought-to-be' evaluations can also be used to justify the imposition of certain normative consequences, adverse or positive, incurred by the agent, despite the fact that 'ought' in the sense of ought-to-be' does not imply can.[33]

2.5.2 Voluntariness and responsibility

But why do we require voluntariness for responsibility and for the range of normative consequence the imposition of which does presuppose the agent's responsibility? Why is it assumed that for ascriptions of responsibility 'ought implies can' remains true? Here we reach the first question raised at the beginning of this section.

Note that it would be easier to answer that question, if one could establish that the imposition of any adverse or positive normative consequence presupposed voluntariness. However, the foregoing discussion of 'ought-to-be' type appraisals suggests that this idea is at least partly mistaken. What I have tried to show in this section thus far is that it can even be appropriate to impose certain normative consequences on the basis of what or how the agent *is*, even if there is nothing she can do anything about what or how she is. If that is correct, then it will not be possible to argue that voluntariness is a requisite condition because the imposition of adverse normative consequences can only be fair provided that the agent could have avoided incurring those normative consequences through his voluntary action.

So why assume that blame or praise-entailing responsibility requires voluntariness? The intuition appears to me to be unassailable. Therefore, I

[32] Complications arise here from the fact that, arguably, (actual) consent impacts on requirements of fairness. In other words, it could be objected here that what the examples in this paragraph show is not that it is fair to impose certain normative consequences on the basis of 'ought-to-be' type appraisals but rather that it is fair impose certain normative consequences provided that there is prior consent to being appraised in a certain way. But even if this is correct, it still remains true that various qualities of the agent can become normatively relevant (via consent), even though there is nothing the agent can do about them.

[33] The claim that blameless agents may shoulder adverse normative consequences draws further support from an important group of cases which involves agents acting in 'limited environments of choice' such as moral dilemmas. The present discussion of 'ought-to-be' type appraisals reinforces our intuition in those cases that the imposition of adverse or positive normative consequences does not require the agent's blameworthiness or praiseworthiness. Just as it seems in certain cases appropriate to impose adverse or positive normative consequences for what or how the agent is, it may well be appropriate too to impose adverse or positive normative consequences for the unavoidable consequences of what the agent does even if the action is overall justified.

would like to defend the view that, although the imposition of various adverse normative consequences can be appropriate even if the voluntariness-condition is not met, ascribing responsibility and the normative consequences which presuppose the agent's responsibility can be appropriate if and only if the action was voluntary.

Let us imagine a community, call it C_1. C_1 has a normative system, by which it evaluates and regulates the behaviour of its members, a system that appears to be quite different from our 'moral practices'. C_1 is a purely 'idealist' culture in the sense that members of C_1 ignore the question of voluntariness entirely.[34] They never rely on 'ought-to-do' type evaluations to assess each other's (or their own) actions and characteristics. What matters is always the ideal norm, never the exigencies of the given situation. In C_1, if you break the norm–never mind what led you to do so–disapprobation will be considered appropriate and possibly further sanctions too. If a promise is broken, members of C_1 decide the question whether or not breaking the promise was objectionable and whether the promise-breaker ought to be disciplined for her misconduct solely on the basis of the pertinent norm. It does not matter why the promise was broken–not even that the promise-breaker was physically coerced into breaking the promise, hypnotized or was simply ignorant of the fact a certain action of his would amounted to breaking his promise–the breaking of the promise automatically entails criticism and various concomitant sanctions. By the same token, in C_1 it does not matter how and why you have come to entertain a norm-offending belief, emotion or instantiate a norm-offending characteristics, the offence against the norm is sufficient to invite disapprobation.[35]

I do not think that the normative system of C_1 is impossible or rests on some kind of conceptual confusion. Nor do I think, on the other hand, that it is immune to criticism. But that criticism cannot simply consist in

[34] It is usually said that *shame*, as opposed to guilt, has "a looser tie to the voluntary", Baier, 'Moralism and Cruelty: Reflections on Hume and Kant,' 279. See also Gibbard, *Wise Choices, Apt Feelings: A Theory of Normative Judgement,* 141-7. The claim is that we tend to experience shame (and expect others to be ashamed) even for things about ourselves that are not brought about by our voluntary actions, e.g. we may feel shame on account of our long-deceased forefathers or fellow nationals. To put the matter in terms of the foregoing discussion, shame is ascribed to a large extent on the basis of 'ought-to-be' type appraisals. Apparently, in many cultures shame plays a much more central role than in Western societies. Accordingly, our imagined C_1 could perhaps be called a pure shame culture.

[35] The norms valid in C_1 can be quite complex and structured. They do not have to state that it is always wrong to break a promise. For instance, to stick to the above example, it may be that breaking a certain promise C_1 in amounts to a very courageous deed in which case no moral disapprobation is called for. Has anything like C_1 ever existed? A community in which a raped woman will be stoned to death if she gets pregnant would perhaps be much like C_1. But, to repeat, there can be a complex set of norms in place in a C_1-type culture so that such horrifyingly cruel measures are never resorted to (despite the absence of the voluntariness-condition).

branding the norms of C_1 as immoral. What needs to be shown is that a system of norms which ascribe responsibility and the concomitant range of normative consequences for voluntary actions only has some aspects to it which are more attractive in some sense. I believe what militates against the normative system of C_1 is the consideration that it does not give members of C_1 a chance to adapt their behaviour to meeting C_1's norms, whatever these may be. The worry is that disapprobation or approbation and the imposition of the concomitant sanctions will be entirely independent from what members of C_1 were and were not in a position to do in a given situation.

What is wrong with C_1 not taking the voluntariness-condition into account, therefore, is that it treats members of C_1 *unfairly.* This is because ascriptions of responsibility and the imposition of the concomitant normative consequences in C_1 do not track an important feature of the relevant situation, namely whether or not people had a genuine opportunity to adjust their behaviour to comply with the norms of C_1. Because of this punishment, and the imposition of similar normative consequences, runs the risk of being *undeserved* in C_1.

But when say that C_1 is unfair are we not just arguing from within our system of norms, call it C_2, where voluntariness is necessary for responsibility and the concomitant normative consequences? Are we merely being chauvinistic? I do not think so. The reason is that C_1 misses a crucial aspect of agency, namely that agents are persons. When people act on norms they do so not because norms cause them to do so but because they regard norms as normative, i.e. as a source of reasons which figure in their deliberations about what to do. The shortcoming of C_1 is that it does not recognize that the norm-subjects are persons who are capable of deliberating about what they intend to do. By eschewing the voluntariness-condition C_1 regards every norm as an external constraint which people may or may not be able to meet as chance would have it. It fails to admit the possibility that norms can influence people's behaviour by guiding people's deliberations when choosing among non-compossible options for action.

I believe that by doing so C_1 remains unresponsive to something that is a constitutive aspect of personhood and because it is constitutive it is also something we value about people. It is this value that is represented by a concept of responsibility which assumes the voluntariness-condition to be necessary for attributions of responsibility and the imposition of normative consequences to be justifiable.

It may be objected to this defense of the voluntariness-condition that ascribing vicarious responsibility is a common practise. One attempted defense of vicarious responsibility could be that it is justifiably employed for the purposes of deterrence and encouragement. That kind of justification, however, is misguided. Applying various sanctions, including overt blame and reproach, exclusively for the sake of deterrence or encouragement may on occasion be justified, but even on those occasions the justification will have nothing to do with the agent being responsible. Rather, it will be said that *despite the fact* that the agent is not blameworthy/praiseworthy it is right to encourage (or discourage) certain forms of behaviour for one reason or another.

Some moral philosophers do not seem to believe that ascriptions of responsibility and the concomitant normative consequences require voluntariness. As already mentioned, Robert Adams has argued that it may be justified to ascribe responsibility for states of mind that are not in our voluntary control. I am not quite certain–despite Adams's occasional insistence to the contrary–that he really does separate the issue of justifying that one is responsible ('X is responsible') from the issue of justifying overt expressions of that judgement ('holding X responsible'). Consider this passage for instance: "The appropriate purpose of reproach, and of judgements of blame, directed at others or at oneself, is not to crush us but to lead us to repentance, and to acknowledge moral realities."[36] If that is the case, his claim that there are involuntary sins can be taken to mean that far fewer normative consequences imposed on the agent in response to his action require that the agent be responsible, then is commonly thought.

At the same time, Adams admits that the imposition of some normative consequences–he mentions punishment in particular–are only appropriate as responses to voluntary actions.[37] But if voluntariness does not matter for (overt) blaming, then why should it matter for, say, punishment? In fact, I take this distinction made by Adams as an indirect admission of the normative significance of the voluntariness-condition.

Consider the case of Gutless Gilbert again. It was stipulated above that he is a cowardly person but that his having this character failing is not his fault in the sense that there is nothing he can or could ever do about his inborn cowardice. Now let us contrast Gutless Gilbert with Hesitant Henry. There is nothing wrong with Hesitant Henry's character. At least, he is known to have acted courageously on several occasions in the past whenever his vital interests were at stake. In fact, we have all seen him dive into the water to recover his expensive surfing board despite the presence of ravenous sharks and tempestuous weather. But now he refuses to do the same to save a drowning child under the same conditions.[38] It is clear, I think, that even if it is, other things being equal, appropriate to impose certain adverse

[36] Adams, 'Involuntary Sins,' 24.

[37] See ibid., 21.

[38] I sidestep here various complications concerning the notion of character. Should we say in view of his failure to rescue the drowning child that the adequate description of Hesitant Henry's character would not be that he is courageous but rather that he is 'courageous when his own interests are at stake but not when that of others'? Whatever one's view on these issues, I think the main point still stands: we evaluate character and actions differently and in part ascribe different normative consequences on the basis of these different appraisals.

normative consequences on Gutless Gilbert for his unfortunate moral failing, these normative consequences will not be the same as those which would be prepared to impose on Hesitant Henry for his morally objectionable action. I submit that our readiness to treat the two cases differently has largely to do with the fact that our 'ought-to-be' type appraisal of Gutless Gilbert's character does not presuppose voluntariness, whereas our 'ought-to-do' type appraisal of Hesitant Henry's action does.[39] The voluntariness-condition can be crucial in determining what kind of normative consequence we consider to be fair to impose on the agent.

If that is true, then we do indeed distinguish between normative consequences that presuppose voluntariness and those which do not. No doubt some normative consequences, both adverse and positive, can be imposed even in the absence of voluntariness. Yet some normative consequences appear to require responsibility. This shows, on my view, that we indeed attribute great normative significance to the voluntariness-condition.

2.6 The normative consequences of action-ascriptions

In this section, I would like to elaborate the claim made earlier that attributing actions to agents can give rise to normative consequences with or without the agent's responsibility. In the following section, I will go on to discuss the specific reasons for action generated only by ascriptions of responsibility.[40]

First of all, however, I think it is important to say more about the nature of normative judgements in general. The normativity of such judgements is based on the *evaluative* character of these judgements. Consider aesthetic judgements, for instance. Such judgements provide us with reasons because of the appraisal they involve. In other words, the evaluation focuses on a given aesthetic feature and gives us reasons to occupy a certain stance (e.g. admiration) towards the work embodying that feature. Thus evaluations generate practical reasons which are not entailed by mere descriptions.

[39] But what about Gutless Gilbert's *actions?* Suppose that it is Gilbert who refuses to save the drowning child. Suppose, in fact, that Henry and Gilbert are both standing on the shore watching the child drown (while not seeing each other). Would we appraise Gilbert's and Henry's failure to act differently? The answer to that question depends, I think, on our view of the connection between character and actions. If we accept that Gilbert's character trait can exercise such a powerful hold on him that it completely and under all circumstances, including the present situation, prevents him from doing the right deed, whereas the same is not true of Henry, then it seems to me to be right to appraise Henry's failure to act differently from Gilbert's failure to act. Meanwhile, it is still open to us to evaluate Gilbert's character very negatively indeed.

[40] Note again that reasons for emotions will be treated as a subclass of reasons for action.

Such an understanding is, I would argue, also applicable in the area of ascriptive theory. In particular, the Ascription Thesis makes explicit reference to the 'normative consequences' ensuing from the appraisal of the agents' actions. Now it is possible to characterize these 'normative consequences' more closely. I contend that there is a special class of reasons that are consequential upon agents' actions (and a smaller class of special reasons that are consequential upon actions for which agents are responsible). We would not have these reason, unless the agent had acted.

The reasons for action generated by action-attributions may be *pro tanto* reasons for the agent to do certain things in consequence of his action. But action-attributions may also generate reasons for *other* people to do certain things in consequence of the agent's having acted. As to the former category, the agent has reasons to do a number of things. Depending on the nature of the situation, an action-attribution (even without a corresponding ascription of responsibility) may give the agent reasons to apologize for what he has done, compensate the victim(s) of his action, seek to justify his action or simply to explain why he has done what he has done, as well as many other things. On the positive side, an action-attribution may give the agent *pro tanto* reasons to lay claim to certain benefits.

But again, it is not only the agent who can have resultant reasons for action in consequence of his action. Other people may have reasons to do certain things in consequence of the action as well. Of particular prominence is the position of those directly affected by the agent's action. For instance, a person suffering harm as a result of the agent's action may often expect to be compensated by the agent, may have the right to demand an apology or at least an explanation, and may have *pro tanto* reasons to express overt criticism of the agent's action. Nor is it the case that only those directly affected by the action have resultant reasons as a consequence of the action. Unaffected bystanders can also call upon the agent to provide reparation for his action and they may also have reasons to criticize the agent for his action. Conversely, they may also be expected to reward the agent for his action, even if they have not directly benefited from the action.

One important type of reason for action we have in response to agency, with or without the agent's responsibility, is what I would like to refer to as the 'justificatory pressure' on the agent. This is itself a normative consequence of the action. Whether or not the agent was responsible, he is expected to *try to* justify what he did and why he did it.[41] As the Ascription Thesis stipulates, if the action was wrong and the demand for justifications cannot be met, then the agent will be blameworthy. But, as we have seen,

[41] I owe this point to John Gardner. As he puts it: "[...]so far as criminal lawyers are concerned, the acquisition of a moral duty to offer some justification for what one did is *itself* a normative consequence of doing it", Gardner, 'In Defence of Defences,' 256. The remainder of the passage makes it clear that Gardner does not limit the scope of this 'duty to justify' only to legal contexts.

even if the demand for justifications can be met, the agent may incur normative consequences in response to his action. Whichever will be the case, the justificatory pressure will be present. In sum, whether or not the agent was responsible, the agent has reasons to offer some justification for what he did and why he did it, and other things being equal, others may demand such justifications. Moreover, I think that the agent can be expected to justify his course of action even if what he did was right. He will be expected to explain that he did right not just accidentally but for the right reasons. All the more so if he lays claim to further benefits, e.g. preferential treatment of some kind or overt praise, as a consequence of his action.

But justificatory pressure is not the only normative consequence of action which is justifiable independently of the agent's responsibility. There is a good case to be made that (i) there is a resultant reason to compensate those who suffered a loss as a result of one's action *irrespective of whether or not the action was justified all-things-considered;* and (ii) that specifically the agent who acted has this reason.[42] In fact, several authors have defended the view that the resultant reasons in this case are mandatory for the agent who acted, i.e. the agent incurs a *duty* to compensate the 'victim' of his action.[43] Note that this is a stronger claim than merely saying that the agent has a non-mandatory reason (rather than a clear-cut duty) to provide some form of reparation to those who suffered a loss as a consequence of his action even if his action was justified overall. Whether or not the stronger claim is also true, the nearly universal consensus with regard to the first, weaker claim indicates that compensation constitutes another instance of justifiably incurred normative consequences with or without responsibility.

In addition, action-attributions in general provide reasons for characteristic emotional responses too. Such emotional reactions form an important subclass of resultant reasons generated by the agent's action. Reasons for entertaining such emotions can be had not only by the agent who acted but also by others, whether or not they were directly affected by the action. There will be more discussion of these reactive emotions and their relationship to responsibility-ascriptions in subsequent chapters.[44] Let me only state at this point that these emotions are individuated by their cognitive content. For instance, guilt is singled out by the agent's belief that she has done wrong, and one's anger by the belief that one has been wronged or one's right violated.

It is important to see that certain reactive emotions qualify as adequate normative consequences in response to action with or without the agent's

[42] Clauses (i) and (ii) are both needed because it may be true that the costs of compensation are to be shared more widely.

[43] Feinberg, *Rights, Justice, and the Bounds of Liberty: Essays in Social Philosophy* and Thomson, 'Imposing Risks.' Cf. Zimmerman's 'Rights, Compensation, and Culpability' for the opposing view.

[44] Especially Chapter 4.

responsibility. Guilt, the example used above, is typically taken to presuppose blameworthiness and is hence singled out as a specific emotional reaction to actions for which the agent is responsible.[45] Other reactive emotions, however, including negative ones, such as *shame, compunction, regret, indignation,* etc. and positive ones too, such as *commendation, adoration, exaltation,* etc. are not connected in the same way to the agent's responsibility.[46] Such reactive emotions appear justified in response to something the agent has done even when the agent is not responsible for that action. In other words, the entertaining of emotions belonging to the latter category can be justified even when the agent could not have done otherwise, i.e. when the voluntariness condition is not met. Thus it may very well be appropriate for Petty Peter to feel ashamed for having taken change out of his friend's purse even if he acted under compulsion or duress.

The issue of reactive emotions leads to the other fundamental tenet of the conception I would like to defend in this work. This is the thesis that it is the cognitive content of our reactions to agency (with or without responsibility) which is the source of the normativity of these reactions. In other words, the claim is that action-attributions and ascriptions of responsibility are normative *because* they are judgements, i.e. because they entail certain beliefs about agents and the circumstances of their actions.

Clearly, our reactions to agency (with or without responsibility) is intimately bound up with emotions, such as anger, guilt, indignation, etc. However, there is an influential group of theories including emotivism, norm-expressivism and Strawsonian naturalism, which I believe tend to overemphasize the importance of this connection.[47] In doing so they obscure the crucial point that it is the cognitive content, i.e. the relevant beliefs, which give these responses their 'normative edge'. It is the cognitive content that helps us to set ourselves at the required critical distance from emotional responses which are unjustifiable or irrational. Because of its insistence on the latter point the account to be defended in this work will be a *cognitivist* one.[48]

[45] What is the emotional counterpart of guilt on the positive side? The difficulty in identifying this emotion indicates that the normative consequences of action-attribution may not divide into two neatly symmetrical categories, one negative and one positive, of roughly equal scope and significance.

[46] More on the classification and normative relevance of reactive emotions in Chapters 3 and 4.

[47] I will discuss Strawsonian naturalism, the most sophisticated and insightful of these theories, in Chapter 4.

[48] For a detailed discussion of the implications of the cognitivist label including the metaethical ones, see Chapter 5.

2.7 The normative consequences of responsibility-ascriptions

Different kinds of normative judgements give rise to different reasons for action. Specifically, I would like to argue that if the agent is judged to be responsible for his action, then that appraisal will produce reasons for action *over and above* those generated by an action-attribution without responsibility.

My principal claim here is that the imposition of certain normative consequences is justified if and only if the agent is responsible his action. Familiar examples of these special normative consequences include certain reactive sentiments as well as certain courses of action to be taken by the agent or others. I do not think it is necessary to provide here a complete list of all normative consequences that presuppose the agent's responsibility. In order to highlight the special normative force of responsibility-ascriptions, it is enough to show that there are at least some normative consequences which we only have reason to impose if the agent is responsible, but not otherwise.

I believe that such a complete list cannot be provided partly because for each type of such normative consequence a separate argument is required to show that it indeed requires responsibility. This is, for instance, what I will undertake to do for some reactive sentiments in Chapter 4. Anticipating that discussion let me provide only one example at this point: It seems to me that the agent does not have a reason to feel *guilty* for his action unless he is blameworthy for it.[49]

Yet another type of normative consequence that presupposes the agent's responsibility is *punishment.* This crucial example has already come up above. It is worth adding here that the case of punishment is complicated since several conditions need to be satisfied simultaneously in order for the agent to be liable to punishment at the first place (for instance, the wrong done has to be grave enough to merit the harsh treatment which the wrongdoer is exposed to). A further distinction of considerable significance is that drawn between the justifiability of carrying out *individual* acts of punishment (within or without the legal domain) as opposed to the justifiability of a whole *system* of punishment.[50] For these reasons, the agent's responsibility is a necessary but not a sufficient condition of his liability to punishment and *a fortiori* a necessary but not a sufficient condition of the actual execution of punishment.

At the same time, I think we can ignore these complications for the moment and argue that *one* of the necessary conditions of liability to pun-

[49] Of course, the agent may have the *mistaken* belief that he is blameworthy. In that case, although it is rational for him to feel guilty, he does not have a reason to feel guilty because he is not in fact blameworthy.

[50] On this point, see Rawls, 'Two Concepts of Rules,' esp. 21-9 and Scanlon, *The Difficulty of Tolerance,* 220.

ishment is indeed the agent's responsibility as blameworthiness. A principal reason for this is that punishment serves important normative purposes which is only possible if punishment indeed presupposes responsibility. One such normative purpose is to express disapproval and reprobation in response to the violation of the law by the punishing authority, an expression which is made on behalf of the entire law-abiding community.[51] Another such normative purpose is, in Scanlon's words, the "affirmation of the victims' sense of having been wronged".[52] I think these are two distinct, though not unrelated, symbolic functions of punishment.

Neither of these normative purposes would make sense if the voluntariness-condition was not met in the case of the action which is to be punished.[53] What is disapproved of by law-makers and the rest of the community is the wrongdoer's *deliberate* perpetration of the criminal act or at least his culpable negligence in bringing it about. What the victims resent and want to have acknowledged through the act of punishment is the criminal's *wilful* offence against them (or his wilful negligence). This sort of resentment is different from whatever they may feel about the actual loss (material or otherwise) they have suffered. Victims of crime may in many cases also legitimately lay claim to be compensated for the loss they have suffered. This, however, is a separate issue which is shown, among others, by the fact that many victims would not rest satisfied with a mere compensation without punishment of the perpetrators no matter how high the payment may be. Moreover, such compensation, even if accepted by the victims themselves, would not serve the other symbolic end of imposing punishment, i.e. expressing resentment of the community at the criminal's deliberate disregard of the community's laws.

As argued above, voluntary all-things-considered wrongdoing is blameworthy. From this it follows that in order for the normative purposes of punishment to make sense, the agent must necessarily be blameworthy for the act punished. If so, then the example of punishment can illustrate how appraising the agent as responsible can generate reasons to act in consequence of her action.[54]

[51] See Feinberg, *Doing and Deserving,* 96.

[52] Scanlon, *The Difficulty of Tolerance,* 221.

[53] These reasons cut across the debate between retributivism and rival theories of punishment. It seems to me that all theories of punishment agree that punishment is to serve these normative purposes (in addition to whatever else it may serve). That suffices to show that punishment indeed presupposes blameworthiness. On retributivist accounts, punishment is justified because wrongdoers deserve to suffer. Here the requirement that the agent be blameworthy is even more crucial to justify punishment.

[54] The reason-based perspective outlined here throws light on what seems to be a disturbing asymmetry in the structure of the normative consequences of action-attribution which I have already indicated in connection with the difficulty of finding a positive counterpart to the reactive sentiment of guilt. In general, this asymmetry has two related aspects. First, it seems that reasons for action resulting from blameworthy actions are much more numerous and more urgent than resultant reasons of praiseworthy actions.

This concludes my discussion of the conceptual framework in which I propose to locate responsibility. I will rely on these definitions in the exposition of rival theories of responsibility which will take up the next three chapters.

2.8 Conclusion

In the previous section, I sought to show that for us to have *pro tanto* reasons to impose a distinct range of normative consequences the agent's responsibility is required. Further, I have argued that the agent's responsibility is required because we believe it to be a necessary condition for the justifiability of the imposition of these normative consequences that the action was voluntary. Earlier on I have argued that voluntariness has special normative significance–that is to say, we have special *pro tanto* reasons for action in response to action if (and only if) the voluntariness-condition has been met–because only if the action was voluntary did the agent have a genuine opportunity to act as reason required, i.e. to adjust his behaviour to comply with whatever norms were applicable in the given situation.

But why do we care about being judged to be agents who voluntarily acted as reason required? Why do we predicate certain normative consequences on this feature of agency? My question here is not why reasons have normative force on agents. Rather, the question that needs to be answered is why we attach special significance to the question how agents voluntarily respond to this normative force when deliberating how to act? We could of course say that it matters to us because reasons themselves matter. This is the structure of the consequentialist conception of responsibility, for example. According to that conception, it matters to us how agents respond to what they had (most) reason to do in a given situation only because we want to maximize the frequency of actions in the future which are in accordance with what there is (most) reason to do. But, as I will argue in the next chapter, that conception is unconvincing. Agency matters to us not only insofar as it is capable of impacting on what happens in the world in the future.

I believe, however, that a similar structure can detected in many non-consequentialist conceptions of responsibility as well (while the forward-looking orientation of consequentialism is of course done away with). The basic idea of these is that agency matters to us because accepting moral

Second, the resultant reasons of praiseworthy actions are shouldered almost exclusively by others, i.e. not by the agent who acted rightly. This asymmetry has led some to accuse our 'morality system', marked by such an understanding of responsibility, of being vindictive and excessively punitive. See for example Baier, 'Moralism and Cruelty: Reflections on Hume and Kant' and Wertheimer, 'Constraining Condemning.' This asymmetry may lead one to question my approach which treats matters related to praise and blame as two sides of the same coin, predicated on the same normative and justificatory conditions.

requirements commits us to wanting those moral requirements met by everyone who is capable of doing so in the past as well as in the future. We do not want to be and want others to be people who Φ-ed, if Φ-ing is wrong, and we do not want this *because Φ-ing is wrong.* The normative force of responsibility-ascriptions is based on the fact that ascribing responsibility to an agent for a right or wrong action is itself evaluative of the agent since it was the agent who brought moral rightness or wrongness into the world, so to speak, by complying or not complying with moral requirements. This is why, it is argued, determining the conditions of ascribability cannot be independent from our acceptance of moral norms or principles. And arguably, this is why we have special reasons for action if actions do not meet those requirements: our justified interest in the requirements themselves gives us *pro tanto* reasons to impose special normative consequences on those who could have met those requirements, but chose not to.

I would like to essay a different route towards grounding the normativity of responsibility-ascriptions. The proposal to be developed in Chapter 6, which I will call the Value Thesis, is that our interest in the ability of persons to voluntarily act (or not act) as reason requires is not derived from (but of course not incompatible with) our commitment to moral requirements. We value this ability of persons because it is constitutive of how we relate to persons (including ourselves). In short, responsibility is a valuable aspect of personhood independent from whatever other moral and non-moral values we may want to recognize.

Whatever its merits, the Value Thesis is grounded in a cognitivist conception of responsibility which regards an ascription of responsibility as a normative judgement concerning the agent and the circumstances of the action. So before the Value Thesis can be developed, we need to get clearer about what claims a cognitivist theory of responsibility really subscribes to and how these claims can be defended against the competition. For this reason in the next three chapters I will review what are probably the most influential theories of moral responsibility. In Chapter 3 I will focus on consequentialist accounts of responsibility. This will serve, among others, to better articulate and defend the distinction I have appealed to already, namely the distinction between 'being responsible' and 'holding responsible'. Through a critical discussion of Strawsonian 'naturalism', Chapter 4 will seek to defend the claims that responsibility-ascriptions are judgements and that we have reasons to act on these judgements (in response to agents) only if these judgements are themselves justifiable. Chapter 5 will further elaborate the criticisms made above of 'objectivist' views of responsibility which hold that being responsible is an objective property of agents independent from our normative or value-commitments. Doing so will lend support to the argument against skeptics who maintain that responsibility-ascriptions are unjustifiable because we have no access to the objective properties which could make responsibility-ascriptions true. Finally, in Chapter 6 I will put forward the Value Thesis and try to defend it against objections from those who think that our commitment to responsibility is dependent on our commitment to morality.

Chapter 3

Consequentialist Theories of Moral Responsibility

3.1 Introduction

In this and the following two chapters, I am going to discuss theories of moral responsibility which differ in some crucial respects from the conception to be defended in this work. Criticisms of these alternative approaches may help to establish the significance of the two basic distinctions which allow us to identify the most important philosophical differences among various theories of responsibility. These criticisms should also help to position the cognitivist theory to be defended in this work relative to available alternatives.

The first of these distinctions is between 'being responsible' and 'holding responsible'.[1] The second distinction is between different understandings of what is essentially involved in an ascription of responsibility: a judgement of responsibility as opposed to a reactive emotion.[2] The view defended in this work is that a cognitivist theory must give priority to the first element in each of these pairs.[3] Such a theory claims that 'being responsible' is prior to 'holding responsible' and that cognitive content is prior to the emotional reactions provoked by the action. The combination of these two claims is the Priority Thesis.

The cognitivist theory advocated in this work holds that the Priority Thesis is necessary to justify and explain the reason-giving force of responsibility-ascriptions. If that is true, then the issue of priority has both *justificatory* and *descriptive* relevance. On the one hand, the justifiability of different manifest responses to action (holding responsible) depends on whether the agent is or is not responsible. Also, the justifiability of reac-

[1] This distinction was introduced in Chapter 2, Section 2.2. See esp. p. 14f.

[2] For this distinction, see Chapter 2, Section 2.6, esp. p. 36.

[3] For a discussion of the metaethical implications of the cognitivist label, see Chapter 5.

tive emotions depends on the cognitive content of responsibility-ascriptions. On the other hand, the issue of priority has *descriptive* significance as well. That is to say, the claim is not only that we *should* give priority to the judgement that the agent is responsible, but also that we typically do so in our imputations of responsibility in everyday practice. If that is true, the Priority Thesis is needed not only to justify ascriptions of responsibility, but also to explain how the psychological mechanisms function which produce such ascriptions. Or so I will argue.

The main weakness of the two different types of theory to be discussed in this and the following chapter lies in overlooking or deliberately sidestepping one or both of these distinctions (my objections to the third type of theory, to be discussed in Chapter 5, will be of a different nature). Consequentialist theories of moral responsibility, which will be addressed in the present chapter, give priority to judgements of responsibility over reactive emotions, but confuse holding responsible with being responsible. The second type of theory, based on Peter Strawson's seminal work on moral responsibility,[4] can be charged with confusing what it is to hold someone responsible and what it is to be responsible and also with being ambiguous at best about the priority of cognitive content over reactive emotions.

At the same time, all of these theories contain important insights regarding the nature of responsibility and the actual practice of responsibility-ascriptions. These insights are to be accommodated by a cognitivist theory of responsibility too. Therefore, these accounts deserve a fair hearing which they do not always receive. This is particularly true of the consequentialist theory of moral responsibility, "the position everyone loves to hate",[5] to which I now turn.

3.2 Three theses

The consequentialist conception of moral responsibility consists of a small number of concise, snugly-fitting theses.[6] Its parsimony is in fact one of the chief virtues of this theory. I will begin by summarizing these theses before going on to discuss the theory and the main objections to it in more detail. After evaluating these criticisms together with ever more sophisticated reformulations of the consequentialist position, which have been made in response to the critics, I would like to consider in closing what, if any-

[4] See Strawson, 'Freedom and Resentment.'

[5] Arneson, 'The Smart Theory of Moral Responsibility and Desert,' 233.

[6] This conception is spelled out in Smart, 'Freewill, Praise, and Blame,' Schlick, 'When is a Man Responsible?' (originally Chapter 7 of his *Problems of Ethics)*, Nowell-Smith, 'Freewill and Moral Responsibility' as well as in his *Ethics.* For more recent examples, see Dennett, *Elbow Room* (esp. Chapters 3-5) and Arneson's more innovative 'The Smart Theory of Moral Responsibility and Desert.'

thing, can be salvaged from the the consequentialist theory by an alternative cognitivist conception of moral responsibility.

The *first* thesis identifies the aim or rationale of responsibility-ascriptions. It is argued that these ascriptions are invariably forward-looking having the principal function of deterring or encouraging certain actions in the future. This the Forward-Looking Thesis. According to this thesis, ascription of responsibility is tied up with manifest responses to action ranging from overt statements of blame/praise to the imposition of punishment or the granting of rewards. Judging that A is responsible for an action serves to identify the right subject at whom such a response is to be addressed. The overt statement of responsibility itself can have a deterrent or encouraging effect on A himself or others. But more frequently holding someone responsible goes beyond a mere statement of responsibility also involving various forms of treatment of the agent such as punishment, penalty, shaming or the handing out of rewards and other signs of positive recognition.[7] On this view, therefore, imputations of moral responsibility are made with a firmly forward-looking purpose in mind, namely to reduce the frequency of the prospective occurrence of certain actions and increase that of others. It follows that the practice of attributing moral responsibility is seen as a more or less institutionalized form of social control.[8]

The *second* thesis holds that responsibility-undermining conditions are to be recognized when the ascription of responsibility cannot fulfil its purpose, i.e. when the imputation of responsibility cannot *influence* the agent in the appropriate manner. This is the Influenceability Thesis.[9] Such is the case when standards excuses and exemptions apply, that is when the agent acts under irresistible compulsion, duress or necessity or when his cognitive and/or emotional capacities are seriously impaired (as in the mentally handicapped) or not-yet-developed (as in children) or when he is non-culpably ignorant of the consequences his action. In all of these cases the application of positive or negative sanctions seems pointless because either the agent is incapable of altering his behaviour (as in cases of impairment or handicap) or would not have behaved in the way he did if only he had been given

[7] See Smart, 'Freewill, Praise, and Blame,' 304-5; Nowell-Smith, 'Freewill and Moral Responsibility,' 56-8; Dennett, *Elbow Room,* 158; Schlick, 'When is a Man Responsible?,' 60-1.

[8] In light of this, it is hardly surprising that consequentialists frequently treat *punishment* as the paradigmatic type of normative consequence of an ascription of responsibility. See most poignantly in Schlick, 'When is a Man Responsible?,' 61: "Hence the question regarding responsibility is the question: Who, in a given case, is to be punished?". Dennett explains this emphasis on punishment by pointing out that punishment is the "most explicit (public, codified, instituted) response", Dennett, *Elbow Room,* 158.

[9] For adopting this thesis, consequentialist theories are sometimes referred to as "influenceability theories", see esp. Scanlon, 'The Significance of Choice,' 159. For the same reason, elsewhere they are said to represent the "economy of threats" approach, see Wallace, *Responsibility and the Moral Sentiments,* 54-61. Wallace adopts the expression "economy of threats" from Hart.

a chance (as in cases of duress, compulsion and necessity). But since on these grounds any manifest response is pointless, necessarily the judgement of responsibility will lose its point too: "We do not charge an insane person with responsibility, for the very reason that he offers no unified point for the application of a motive".[10]

It is worth noting that the order of these two theses is sometimes reversed. That is, the truth of the first thesis can be argued for on the basis of the intuitive plausibility of the second. Thus one may want to begin by thinking about the considerations that typically lead one to recognize certain excuses as valid and then proceed to enquire what those considerations reveal about the underlying rationale or purpose of responsibility-ascriptions. After all, to all appearances it is a good reason for recognizing certain conditions as excusing the agent from responsibility that in situations where these conditions obtain the ascription of responsibility and the concomitant imposition of normative consequences can have no effect on people's behaviour. If an ascription of responsibility is bound to be ineffective, why bother with the question of the agent's moral responsibility any further? But the intuitive appeal of that consideration lends plausibility to the Forward-Looking Thesis according to which, as we have seen, ascriptions of moral responsibility are essentially driven by a forward-looking concern, namely to "spur" people (Smart's expression) to act as is morally required of them.

The *third* thesis is about the kind of freedom agents can have if determinism is true. It is argued that even if determinism is true agents can be sufficiently free to be held morally responsible. This is the Compatibilist Thesis. This thesis rests on the thought that the freedom required for moral responsibility is the freedom to be able to act otherwise so long as one chooses (or desires or wills) otherwise. There is general agreement among consequentialists that this sort of freedom (as opposed to libertarian freedom) is *not* made impossible by determinism. What's more, the truth of determinism may in fact be positively required for the ability to do otherwise.[11]

Once again one finds an encouraging fit between the Compatibilist Thesis and the Forward-Looking and Influenceability Theses. Thus note that the necessary condition of freedom stipulated in the Compatibilist Thesis is *not*

[10] Schlick, 'When is a Man Responsible?,' 61.

[11] See ibid., 62: "The absence of the external power expresses itself in the well-known feeling (usually considered characteristic of the consciousness of freedom) that one could have acted otherwise... It is of course obvious that I should have acted differently had I willed something else; but the feeling never says that I could also have willed something else, even though this is true, if, that is, other motives had been present." See also Smart, 'Freewill, Praise, and Blame,' 299: "In moral contexts the conditions that are of most importance are 'if he had chosen', 'if he had tried', 'if he had wanted to'." For Dennett this is the only kind of freedom "we care about", see his *Elbow Room,* 139. The claim that the truth of determinism is positively required for moral responsibility is made among others by Smart in his 'Freewill, Praise, and Blame,' 302-3.

met when valid responsibility-undermining conditions obtain, i.e. when the ascription of moral responsibility cannot influence future behaviour. Moreover, it seems that the truth of determinism makes existing strategies (e.g. punishment) no less potent in influencing people's prospective behaviour.

Further, by adding compatibilism to the mix it becomes clear *how* responsibility-ascriptions and the concomitant normative consequences are supposed to influence future behaviour: they are to impact on the agent's motives and desires as well as, over the long haul, on his character. As one author puts it: "Rewards and punishments are means of varying the causal antecedents of action so that those we desire will occur and those we wish to prevent will not occur. Cleverness and industriousness are both valuable characteristics; the latter is called a 'moral' one and the former not, because we know from experience that the former cannot be induced by means of praise and blame, while the latter can."[12]

3.3 The 'elegance' of consequentialism

Already at this point the principal worry about the consequentialist account becomes palpable. It seems to follow from the above characterization of effective ways of influencing people's behaviour that on this account the agent does not need to know *why* he is being encouraged or deterred by means of punishment or otherwise as long as the sanction is effective in putting him on the right course by altering his desires, motivational set or character. But if that is so, then it is questionable in what sense the ascription of responsibility and the concomitant normative consequences are addressed at the agent at all. In other words, it looks like punishments and rewards based on ascriptions of responsibility amount to little more on this account than positive or negative stimuli applied in order to control the agent's behaviour. However, before proceeding to criticisms of the consequentialist theory of moral responsibility, it is worth pausing a little to appreciate some of the attractive features of this position.

To begin with the last thesis, it is often said in favour of compatibilism that it can dispense with the metaphysically strenuous assumptions of libertarianism, such as the possibility of contra-causal freedom or self-causation, on which at least some versions of libertarianism are based. Compatibilism has no need of a "a metaphysical deus ex machina",[13] that is, no need to take "recourse to the obscure and panicky metaphysics of libertarianism".[14]

[12] Nowell-Smith, 'Freewill and Moral Responsibility,' 59.

[13] Ibid., 51.

[14] Strawson, 'Freedom and resentment,' 25. At the same time, at least some compatibilists recognize that their position may contrast with ordinary language as well as everyday practices of responsibility and punishment because these reflect deeply-rooted incompatibilist intuitions. On this point, see esp. Nowell-Smith, 'Freewill and moral responsibility,' 45 and Smart, 'Freewill, Praise, and Blame,' 300. The upshot is that com-

But of course this applies to compatibilist theories of a non-consequentialist kind as well, i.e. to those which do not subscribe to the first and second theses.[15] There are nevertheless other appealing features specific to the consequentialist position. Thus consequentialism directs our attention to the central role responsibility-ascriptions play in shaping social interactions as well as in guiding critical self-reflection. The insight here is that these ascriptions and the forms of sanctioning behaviour they give rise to can function as highly effective instruments of controlling one's own and others' behaviour through their capacity to provide the incentives for education, deterrence and encouragement.

Further, this theory is capable of providing a unified account of what holds various responses to action together as a class, namely the forward-looking concern stipulated in the Forward-Looking Thesis. This is a considerable achievement because the behavioral manifestations of these responses–ranging from blaming/praising to punishing/rewarding–may be quite heterogenous indeed, some being verbal, some not, some being directed at others, some at oneself, some being positive, some negative, etc. Most importantly, by providing such an account consequentialism can elegantly explain the normativity of these responses as well: their normative appropriateness is a function of how well they contribute to encouraging/discouraging desirable actions in the future.

In addition, a consequentialist theory of moral responsibility meshes smoothly with a broader consequentialist outlook in morality whether of a classical Utilitarian or more contemporary ilk. That is important because in this way this account can readily answer the question concerning the place of ascriptive theory (dealing with agency, responsibility, etc.) within practical philosophy.[16] Ascriptive theory on this view deals primarily with those causal antecedents of action which are susceptible to external influence (even if determinism is true). In other words, consequentialism can elucidate the moral significance of responsibility, i.e. why morality is inconceivable without a distinct notion of responsibility and agency (a considerable problem for some alternative cognitivist positions as will be seen in later chapters).

patibilism may turn out to be just as revisionist in its stance towards everyday moral practices as libertarianism, possibly even more. Compatibilists who are ready to admit this include Arneson, 'The Smart Theory of Moral Responsibility and Desert,' 238-9 and Wallace, *Responsibility and the Moral Sentiments,* 58, 117 (Wallace is of course not a consequentialist). Further, it will be seen that Smart's (and Arneson's) is not a classical compatibilist position because he seeks to replace judgements of responsibility as they are commonly understood with a different understanding of what judgements of blameworthiness/praiseworthiness *should* be taken to consist in. To anticipate, Smart does not subscribe to the compatibilist view that genuine or 'deep' judgements of moral responsibility are compatible with determinism. This position makes his (and Arneson's) the most radically revisionist of available compatibilist positions. More on this below.

[15] Wallace's *Responsibility and the Moral Sentiments* is one example.

[16] See Chapter 2, Section 2.4 for the division of practical philosophy into value theory, normative theory and ascriptive theory.

Quite simply, we need those notions to more effectively bring about desirable states of affairs by exerting the right kind of pressure through criticism, censure and approval at the right kind of place, i.e. by putting pressure on agents who are capable of furthering this goal at the first place. Or so it is argued.

Moreover, the consequentialist theory can effortlessly explain isomorphisms between law and morality and also the nature of the connection between these two normative domains. It is claimed that both domains serve primarily to prevent unwelcome behaviour and encourage its opposite. Their exists of course a division of labour between them, moral sanctions regulating behaviour beyond the reach of legal prescriptions. Further, moral and legal sanctions mutually reinforce one another working hand-in-hand to ensure compliance and inculcate the right kind of attitudes and commitments in people.[17]

Finally, the position recommends itself by realizing not only theoretical but also moral desiderata. It seems immune to the frequently heard objection that responsibility-attributing practices are driven in a "persecuting spirit" by a righteous and cruel desire to seek vengeance and retribution.[18] Consequentialists dispense with the illusion of fittingness–"a pitiful trinket for a philosopher to wear as a charm against the recognition of his own humanity"[19]–i.e. they can make do without the idea that there could or should be a fitting sanction in response to everyone action, a sanction inherently deserved by those who do wrong (or act rightly). The affirmation of the forward-looking rationale of responsibility-attributions allows them to consistently assume the tolerant and constructive attitude of "bygones are bygones".[20]

3.4 Criticisms of consequentialist theories of moral responsibility

If consequentialism is that attractive why not go for it? Two objections to consequentialism will be considered in this section. They articulate a common worry, namely that the "theory appears to conflate the question of whether moral judgement [of responsibility] is applicable and the question

[17] Note that classical Utilitarians tended to stress the preventive/encouraging function of law regarding criminal and tort law as the foundation of the legal system. See Raz, 'The Functions of Law,' 178.

[18] In line with the enlightened, reformist and socially conscious attitudes dating back to classical Utilitarianism, the crudeness and moral repugnancy of retributive practices is often stressed by those who subscribe to consequentialism about responsibility. See for instance Schlick, 'When is a Man Responsible?,' 60; Nowell-Smith, 'Freewill and Moral Responsibility,' 58, etc.

[19] Strawson, 'Freedom and Resentment,' 24.

[20] Rawls, 'Two Concepts of Rules,' 22.

of whether it should be expressed".[21] However, the case against the consequentialist theory of moral responsibility cannot be settled just by reiterating Scanlon's criticism. This is because for a consequentialist the question whether a judgement of responsibility is applicable is ultimately the same question as whether it is appropriate to express it, whereby appropriateness is assessed in terms of the forward-looking concern. Given that concern, for a consequentialist the question of moral judgement of responsibility just is the question whether that judgement should be expressed. So merely repeating that these are two separate issues would beg the question.

The vulnerability of the consequentialist account can be demonstrated, however, by showing that the forward-looking concern does not motivate ascriptions of moral responsibility and that 'being responsible' enjoys both justificatory (normative) and explanatory (descriptive) priority over 'holding responsible'. Accordingly, the *first* objection is that agents are not always held responsible because holding them responsible is thought to deter or encourage. For several types of sanctioning behaviour the justifiability of sanctioning behaviour appears to depend on the appropriateness of responsibility-judgements themselves.

The consequentialist can try to fend off that objection by extending his list of the ways in which 'holding responsible' can promote the forward-looking concern and by relaxing the requirement that people should always be aware of what the real reasons are for holding one another responsible. But by increasing the number of 'ways of influence' that can be relevant to justification, the consequentialist only makes the worry expressed in the *second* objection the stronger. According to this objection, if the consequentialist chooses to justify sanctioning behaviour exclusively in terms of the forward-looking concern, he will have no satisfactory account at all of the normative role of responsibility-judgements themselves. This is because the consequentialist will no longer be able to explain how and why ascriptions of responsibility are capable of guiding behaviour. If that is true, then the justification of responsibility-judgements itself cannot depend on the forward-looking concern.

The conclusion I will draw from these two objections is that not only 'being responsible' is not dependent on 'holding responsible', but rather the latter is dependent on the former. As will be seen, the dependence is both descriptive and normative. As already mentioned in the introduction in connection with the Priority Thesis, we can only properly explain how practices of holding responsible work by taking the priority-relation in this way and we can only justify these practices by taking the priority-relation in this way. Let me now spell out these two objections in more detail.

[21] Scanlon, 'The Significance of Choice,' 159.

The *first* objection to the consequentialist account of moral responsibility is that the justifiability of manifest responses themselves cannot depend entirely on the potential of any given response to deter and to encourage.

Consider, for example, the case of punishment. The worry here is not the frequently heard objection that consequentialism justifies punishment of the innocent (I will come to that difficulty below). The point is rather that deterrence is at best a necessary condition for the imposition of punishment *even* if it could be somehow ensured within the consequentialist framework that only blameworthy agents are punished.

Consequentialists require for the justifiability of punishment, first, that the punitive sanction be capable of influencing the agent's future behaviour (otherwise he ought to be excused from responsibility), second, that the punishment should effectively further the goal of deterrence, and third, that the agent be guilty (again, let us add this explicitly to avoid the most familiar objection against a more rudimentary form of consequentialism). Note, however, that even if all three of these conditions are met, punishment may still be unjustifiable. This is because punishment is invariably to be meted out by someone, that is to say, imposing punishment is tied to a role or an office.[22] The office is formally circumscribed in the case of legal punishment, but even in the case of informal moral punishment one or more people must by definition assume the role of the punisher (parents, teachers, friends, etc.). For this reason, however, it may well turn out to be the case that even though it would be appropriate to punish the agent in terms of the three conditions just mentioned, one's entitlement to assume the role of the punisher, i.e. to actually impose the punishment, is lacking. A may be liable to punishment (as he meets all three conditions), but B may not be entitled to punish him (or not entitled to do so at a certain time).[23]

There can be various reasons for this. For instance, there may be good grounds to question B's impartiality or his ability to properly assess the situation. In any case, the general point is that punishing (both moral and legal) requires the entitlement to punish. And the fact that the agent is liable to punishment in terms of the three conditions mentioned does not give everyone the authority to hand out punishment. For legal punishment and at least some forms of moral sanction, such authorization is only to be had by a formally appointed body (or person) that initiates a public process according to previously laid down rules subject to customary checks and balances. But some form of entitlement is indispensable even in the most

[22] A point already mentioned in Chapter 2, see p. 37.

[23] Note that the force of this general objection is explicitly recognized by the most sophisticated version of consequentialism to be discussed in Section 3.5 below.

informal instances of punishment. Only a parent or a recognized guardian (if anyone) has the authorization to punish a child in certain ways.[24]

Moreover, even if the punisher is authorized to mete out punishment there may be good reasons, moral as well as non-moral reasons, not to do so. Thus it is conceivable that the agent has, as it is often said, 'already suffered enough', not to mention other reasons for mercy or lenience. Further, it is conceivable that punishment is to be refrained from because its imposition can be expected to significantly increase the future incidence of the act to be punished as in some cases of civil disobedience for instance.

The same point can be made with regard to the imposition of other normative consequences as well. For example, even if an agent has clearly done wrong not everyone is entitled to demand an apology from him for his wrongdoing, not even if his future behaviour could be positively influenced by this demand. These considerations do not constitute a decisive argument against the consequentialist position, but they do show already that the Forward-Looking and Influenceability Theses stipulate what are at best necessary (but not sufficient) conditions for the justifiability of the actual imposition of a distinct range of normative consequences.

The *second* objection goes further. The worry here is that the justifiability of responsibility-attributing judgements is itself not dependent on the forward-looking concern. Moreover, the justifiability of many forms of sanctioning and praising behaviour can also be shown to depend on backward-looking judgements of moral responsibility.

The first thing to note in support of this second criticism of consequentialism is that ascriptions of responsibility can be and often are made *privately.* One may simply lack the opportunity to express one's judgement of responsibility (the addressee is too far away, is no longer alive, there are too many people involved, etc.) or may feel not entitled to do so for reasons indicated above. And yet one can make a judgement of responsibility keeping it to oneself, making a mental note of it, as it were.

Some critics make a lot of this observation.[25] However, it seems to me that the mere fact that responsibility-ascriptions can be made privately has in itself limited purchase on the consequentialist position. This is because consequentialists do not say that ascribing-responsibility is exhausted by the readiness to react to the agent in certain ways. They have a robust notion of what a judgement of moral responsibility is: they are judgements concerning the appropriateness of imposing sanctioning or praising behaviour. It is quite open to consequentialists to say that ascriptions of responsibility made

[24] This applies not only to punishment but also for instance to moral reproach. Only one's closest friends, relatives or partner are authorized (if anyone) to reproach one for some things such as being, say, spendthrift.

[25] See Sher, *In Praise of Blame,* 74, 86 and Wallace, *Responsibility and the Moral Sentiments,* 56.

privately are not merely behavioral dispositions but are nevertheless parasitic on publicly expressed statements of such ascriptions. Privately made ascriptions, on this account, evidence the internalization of norms of social control. They are withheld under the circumstances but they would not be if only the opportunity presented itself to confront the addressee with them.

So consequentialists could indeed argue that the possibility of making responsibility-ascriptions in private does not show that judgements of responsibility are driven by different normative considerations than those regarding how and when these judgements should be expressed.[26] For the consequentialist the crucial point is that we could not imagine such private ascriptions of responsibility without there being a public practice of responsibility-ascriptions in place which practice in turn is driven by the forward-looking concern of social control.

I believe nevertheless that the possibility of making ascriptions of responsibility privately without an intention or opportunity to express them is a symptom of the real difficulty with the consequentialist position which is that consequentialists misconstrue "the real nature of our praise- and blame-related responses".[27] It will be remembered that according to the consequentialist criterion of moral responsibility the agent is morally responsible if and only if it is appropriate to impose normative consequences on him in response to his action: "the question of who is responsible is the question concerning the correct application of the motive".[28] The problem is that this definition misses the essential point that the "attitudinal aspect" of responsibility-ascriptions is "backward-looking and focused on the individual agent who has done something morally wrong".[29] This is shown not only by there being entirely private ascriptions of responsibility, but also by the fact that as a matter of everyday practice, we frequently and typically concern ourselves very seriously with past actions which are not going to re-occur in the future.

It seems, therefore, that the consequentialist theory is wrong in describing judgements of responsibility as motivated by the forward-looking concern. In other words, it is not the case that judgements of responsibility are made depending on whether they would be conducive to the goals of deterrence or encouragement whenever they come to be expressed. Judgements of an agent's responsibility are not utterances of deterrence and encouragement with the volume turned off.

Again, this point has both a descriptive and a normative aspect. At the level of description, it is simply not true that when making judgements of moral responsibility we take into account whether or not a public expression of that judgement would be capable of influencing the agent or others

[26] *Pace* Sher, 74.

[27] Bennett, 'Accountability,' 20.

[28] Schlick, 'When is a Man Responsible?,' 61.

[29] Wallace, *Responsibility and the Moral Sentiments,* 56.

in the right way. At the normative level, the point is that the norms for the appropriateness of the judgement of responsibility do not depend on the norms concerning the appropriateness of public expressions of that judgement. It is perfectly possible that I am justified in blaming the agent for his behaviour privately, but that I am at the same time not justified in openly sanctioning the agent for the very same behaviour (for reasons discussed in connection with the first objection). It may not only be impossible (for lack of opportunity) but also be impermissible to express a wholly justified judgement in any way.

Moreover, the point about the backward-looking attitudinal aspect seems to hold true for sanctions associated with ascriptions of moral responsibility. These also appear to be directed at what the agent has done and not what he is likely to do in the future. Consequentialists often focus single-mindedly on punishment because here it may seem at least conceivable that the institution is maintained while its retaliatory, i.e. backward-looking "attitudinal aspect" is done away with. I shall not argue this point here because there are other forms of sanctioning or praising behaviour where the unfeasibility of such a proposal is more obvious. Thus it will be remembered that reactive sentiments such as anger, indignation, resentment, gratification, admiration, etc. are also normative consequences imposed on the agent in response to his action.[30] There is a good case to be made that construing these moral reactive sentiments in terms of a forward-looking concern lacks any psychological reality, and what's worse, gets the order of justification wrong too. This is because what lies at the core of these responses is the characteristically retrospective belief that the agent has done wrong (or has acted rightly).

We have to take recourse to this 'core belief' both to explain how manifest responses work psychologically and to justify them. As regards the task of explanation, the psychological mechanism of manifest responses is particularly clear in reflexive cases, i.e. when one ascribes responsibility to oneself for something one has done. This kind of self-directed attitude often does not issue in any kind of sanction (against oneself). It does, however, frequently issue in guilt or remorse. The occurrence of such feelings does not seem to depend psychologically on questions of deterrence or encouragement. For instance, the feeling of guilt is triggered by our appraisal of what we have done and not by how we would like to act in the future.

Consequentialists could object here that, whatever the feeling of guilt is triggered by, feeling guilty prevents us or at least discourages us from doing again what we feel guilty about. If that is true, then a feeling of guilt does contribute after all to promoting the forward-looking concern by discourag-

[30] For a discussion of these reactive emotions, see Chapter 2, p. 35 and in more detail Chapter 4.

ing a repetition of the action we feel guilty about. Or so consequentialists can argue.

There is no reason to deny that the feeling of guilt (and other reactive sentiments) can play such a forward-looking role. But guilt, remorse, etc. can play such a forward-looking role by being *motives* and not *instruments* of improvement.[31] They are not somehow instrumentally imposed by the agent on himself in order to discourage himself from acting in certain ways in the future: one does not tell oneself to feel guilty in order to make sure that one will not do the same again in the future. Saying that would fundamentally misdescribe the psychological situation in which feelings such as guilt are entertained. Guilt assails us, feeling guilty is not a matter of decision. But even more importantly, guilt and other reactive sentiments can play an effective motivational role, i.e. feeling them can genuinely discourage from repeating past wrongs, only if they are triggered, as I said above, by an independent appraisal of what one has done, independent that is, from our concern for how best to promote desirable behaviour in the future.

The same holds true for non-reflexive cases as well. It is only an apparent paradox that overt blaming or praising can have instrumental value, i.e. bring about the agent's moral improvement, only if it is deployed for non-instrumental reasons. In the moral domain, sanctioning behaviour can influence the other's behaviour only by causing the blamee ”to believe that the speaker actually has the attitude the behaviour expresses”.[32] Once the blamee realizes that there is a different motivation behind the response, that the response is made merely for the purpose of promoting the forward-looking concern, he will no longer be affected.[33]

By the same token, the forward-looking consideration that the agent ought to do better next time is not necessary to justify the expression of judgements of moral responsibility. The same applies not only to reactive sentiments but also to, say, demanding an apology. For such a demand to be justified it is necessary to believe that the agent has done wrong and it is not necessary (although of course possible) to entertain the expectation that expressing such a demand will prevent him from committing the same kind of act in the future.

The upshot of this objection is that the forward-looking concern does not provide good enough reasons for making ascriptions of responsibility *in just too many cases,* namely in all of those instances when the agent cannot be directly confronted with the judgement (because he is far away, has passed away, will not understand, etc.) or is unlikely to be swayed by the judgement (because the trait giving rise to the action is too deeply entrenched in his character and so on). That is descriptively inadequate: we make attributions

[31] Pace Arneson, ’The Smart Theory of Moral Responsibility and Desert,’ 241, 251.

[32] Sher, *In Praise of Blame,* 74.

[33] Children too will react differently when they begin to see that their parents are not really angry only simulate anger to get them to do something.

of responsibility all the time even when those cases obtain. But it is also normatively misguided since we appear to have very good reasons to make attributions of responsibility even in such cases.

At this point, the consequentialist can seek to strengthen his account by pointing out that there are several ways in which holding an individual responsible can produce good consequences.[34] First, the ascription of moral responsibility and the concomitant imposition of normative consequences can impact on the agent himself by influencing his desires, motivations or by shaping his character. But also, second, the response to the agent, though leaving the agent unaffected, can alter the desires, motivations or character of *others,* who witness that response, and thus potentially increase the likelihood of *them* acting in desirable ways in the future. Third, the mere threat of certain responses can induce people to seek to avoid wrongdoing. Fourth, people may gain satisfaction from seeing wilful wrongdoing not going unpunished and, on the other hand, from seeing virtuous acts praised and rewarded. This satisfaction can itself be counted as a positive consequence but further it can indirectly enhance compliance and cooperation. Finally, ascribing responsibility may give satisfaction to the person who makes the ascription itself. We should not underestimate the satisfaction to be gained from having one's voice heard.

I think none of these points should be denied by opponents of consequentialism. However, they are right to question to what extent the possibility of influencing people's behaviour in these ways really bears on the justifiability of responsibility-ascriptions. Instead of allaying the worries expressed above, the consequentialist's rejoinders make them even more acute. By extending the list of the 'kinds of influence' relevant to the justification of holding people responsible, the consequentialist position makes it look increasingly irrelevant whether the agent who becomes the target of sanctions was in fact responsible or not.[35] The likelihood of putting other agents on the right track will give us good enough reasons to go ahead with the imposition of normative consequences irrespective of the agent's responsibility.

What this highlights is the consequentialist's essential dilemma about the proper source of justification for judgements of moral responsibility. The consequentialist can rightly call attention to the fact that holding an agent responsible can deter or encourage and thus promote the forward-looking concern not just by influencing the agent directly but in other ways too. If he does so, he will certainly improve the *descriptive* accuracy of his account of responsibility-ascriptions as means of social control. After all, each of the 'kinds of influence' listed above is something one can quite realistically take people to have in mind when engaging in sanctioning behaviour (e.g. the

[34] See Arneson, 'The Smart Theory of Moral Responsibility,' 242-3 and 249.

[35] For an example of a radically consequent consequentialist who is prepared to admit this, see Arneson's 'The Smart Theory of Moral Responsibility'. More on his approach below.

teacher may scold a student not in order to influence him, knowing that that is unlikely to happen, but with the purpose of influencing other students in the class). In addition, by extending the list of relevant 'kinds of influence' the consequentialist can lend more credibility to his account of why we *should* engage in sanctioning behaviour. The extension of the list reflects that sanctioning behaviour can improve future prospects in manifold ways and not just by exerting pressure on the agent himself. On the basis of this, the consequentialist can make a more convincing case that the justifiability of such behaviour depends on how effective various responses are for the purposes of deterrence or encouragement.

However, the difficulty on this horn of the dilemma lies in the fact that once the consequentialist goes for the option of extending the list of relevant 'kinds of influence' his position becomes vulnerable to the objection that he severs the link between the judgement of responsibility and the justification of sanctioning behaviour. The problem here is that the utility of sanctioning behaviour now "depends on too many factors other than the nature of the act in question".[36] Once we no longer focus exclusively on influencing the agent himself by means of sanctioning or praising behaviour, it becomes increasingly unclear why the agent's responsibility should matter at all. Or at least, there is no reason why it should always be necessary to justify sanction or praise.

In fact, it seems that, strictly speaking, on this horn of the dilemma, the consequentialist has no time for the notion of moral responsibility at all. The theory, once it begins to broaden the notion of influenceability, cannot explain why sanctioning or praising behaviour should turn on the agent's responsibility. This horn of the dilemma pins the consequentialist because he appears to be forced to give up his commitment to the notion of moral responsibility as a judgement *of* the agent.

On the second horn of the dilemma, the consequentialist continues to hold on to a notion of moral responsibility as a response addressed at the agent. This enables him to maintain that there is a correlation between the moral quality of what the agent has done and the response. But in this case it seems false to say that the judgement of moral responsibility is motivated by the forward-looking concern. It is descriptively false because as a matter of fact we very often have no such concern in mind and normatively false because it may be justified to ascribe responsibility to the agent even if that ascription does not or cannot issue in sanctioning or praising behaviour.

To end this section, I would like to discuss a typical consequentialist strategy to seek a way out of the dilemma here described. This response concedes that the consequentialist theory of moral responsibility misdescribes attributions of responsibility as they are commonly practised in everyday life, but

[36] Scanlon, 'The Significance of Choice,' 160.

insists that these practices are nevertheless morally repugnant and based on theoretically erroneous suppositions. This is Smart's view for example. Smart urges that instead of judging agent's for their actions we should grade them, praising or dispraising agents for their good and bad actions as one would grade an apple or a woman for her nose.[37] It is true that people do not ordinarily think of judgements of moral responsibility in this way, but in light of the consequentialist criticisms of these practices any "dispassionate and clear-headed"[38] person "ought to modify [his] attitudes".[39] In fact, "we should stop judging"[40] agents as morally responsible for their actions, or more precisely, we should think of judgements of moral responsibility as the ascription of the capacity to an agent to change his desires, motivation or character upon being confronted with praise or dispraise.[41]

The principal shortcoming of that recommendation should be already apparent from the foregoing discussion. It is unclear how changes in the agent's desires, motivation or character would be effected by the deflationary notion of responsibility advanced by Smart. If ascriptions of responsibility are mere acts of grading, then it is hard to see how such ascriptions should move agents to mend their ways. So the problem is not simply that we lose our grip on a robust notion of moral responsibility as an evaluative stance towards the agent but–quite apart from the criticisms of the consequentialist approach made above–it is questionable whether *even the forward-looking concern* could be effectively promoted on such a minimalist account of what it is to hold an agent responsible. If your blaming me for what I have done entails, first, that my action was bad (as an apple may be unfit for sale or a woman's nose ugly) and, second, that you think I am such a person who can change his ways in response to such a judgement, I may just not care (as a woman may just not care that Smart finds her nose ugly). I will care if you threaten to punish me but the only reason why I will care is because the punishment may be unpleasant or painful.

In any case, that's just not how it works! Neither descriptively, as regards the psychological mechanisms which motivate us to make ascriptions of responsibility, nor normatively, as regards the criteria of justifiability of responsibility-ascriptions: I care about your blaming me because it entails a negative judgement about the quality of my act and because punishment of it is justified or not depending on that quality.

In response to this, one could seek to improve Smart's consequentialist theory by conceding the point I have just made while insisting that the underlying rationale of responsibility-attributions is nevertheless still to be understood in consequentialist terms. Thus adherents of consequentialism

[37] Smart, 'Freewill, Praise, and Blame,' 303.
[38] Ibid., 305.
[39] Ibid., 291.
[40] Ibid., 306.
[41] Ibid., 304-5.

can maintain that sanctioning and praising behaviour may achieve their purpose *precisely because* those who engage in such behaviour are not actually aware of the underlying forward-looking rationale for such behaviour. For instance, in expressing resentment publicly about your not turning up to our meeting I myself may be thinking that I am scolding you to make you aware that you have done wrong and that's the end of the matter. In fact, however, the reason why my scolding you is justified is because my response may be capable of spurring others to come on time. The suggestion is that "[...] the agent at the time of praising and blaming probably cannot have in mind the thought that she is behaving strategically to induce good consequences. But the conditions that warrant accountability need not be in the mind of someone engaged in accountability practice".[42] Sanctioning and praising responses may be *more* effective, it is said, if one is not aware of what purpose holding one another responsible is really meant to serve, if one continues to believe that such responses are essentially retrospective.[43]

Arneson argues exactly in this spirit that "judging and blaming and shaming in the ways he [Smart] rejects might be valuable instrumental additions to the practice of responsibility".[44] But note that the rationale for taking account of the judgement is only allowed by Arneson as a concession to the psychological reality of how human beings work. The concession is made because it allows for a more effective promotion of the forward-looking concern since it is "quite possible that judgement-inclusive responsibility would outperform responsibility shorn of judgement".[45]

But if not even such a "judgement-inclusive" ascription of responsibility can be expected to influence the agent for the better, then we must simply say that the agent "is not morally responsible for his misdeeds".[46] Such a case is presented in Arneson's thought experiment of the Mafia thug who is only made more irritable and more brutal if criticized or reproached for terrorizing a village. Arneson claims that because the Mafia boss cannot be positively influenced by any condemnation of his behaviour, he is not morally responsible.[47]

[42] Arneson, 'The Smart Theory of Moral Responsibility,' 240n13. See also ibid., 247 for the same point on self-blame: "We should distinguish what holding oneself responsible amounts to and what one should have in mind when reproaching oneself in the course of holding oneself responsible."

[43] And if, furthermore, one continues to believe (erroneously) that the agent was free in a full-fledged, unconditioned libertarian sense of the word 'free'.

[44] Ibid., 239.

[45] Ibid., 242.

[46] Ibid., 248.

[47] As a limiting case, Arneson is prepared to recognize that the mere condemnation of the Mafia boss "can do good by blaming the perpetrator in the hearts [of victims and observers of the thuggery] even if no external expression of such blame is warranted on consequential grounds", ibid., 249. To the extent, but only to the extent, the unexpressed blame "can do good" in this way, the judgement of responsibility is warranted, says Arneson.

But something goes wrong here too. First, how could the *impracticability* of influencing the Mafia boss make the judgement that he is morally responsible for his wrongdoing any less warranted? It is the other way around: the judgement of moral responsibility is warranted but under the circumstances it may not be advisable to express it or impose any normative consequences on him on the basis of that judgement. In fact, the example seems to show once again, contrary to Arneson's intention, that it is misguided to look for consequentialist justifications for ascriptions of responsibility because such ascriptions often do not or cannot issue in manifest responses.

Second, there is Arneson's suggestion that the blissful ignorance of people may, unbeknownst to them, actually further the underlying rationale of responsibility-ascriptions: "[...] praise and blame and the like are natural human reactions; the question is just whether they should be inhibited or encouraged under some circumstances".[48] This proposal harks back of course to a form of Utilitarianism, most closely associated with Sidgwick's work, according to which "full publicity of moral theory to moral agents"[49] may in fact be counterproductive to the promotion of the forward-looking concern. As is often noted, and not only by those with deontological leanings, there is something disturbing about the Utopian élitism of such a proposal and something unpersuasive too about its lack of psychological realism.

Specifically, the current proposal may fail even according to consequentialist standards. Note first of all that the revised view urged by Arneson is even more radical than Smart's. The latter believed that a revised conception of moral judgement can be effective in promoting the goals of deterrence and encouragement. However, not even Smart divorced the practicality of the judgement of moral responsibility as radically from the adequacy of its content (or, which is the same thing, wholly equated adequacy with practicality) as Arneson has done. On Smart's conception, judgements of responsibility, though seriously revised in comparison with our ordinary notions, achieve their purpose because they are true in at least some of the cases. In Arneson's proposal, by contrast, any kind of appeal to adequacy or truth is dispensed with. If on the whole it cannot "do good" to hold the Mafia boss responsible, then he is not responsible.

Again, quite apart from all other pressing worries that radical conclusion makes it even more difficult to see how ascriptions of responsibility should work towards promoting the forward-looking rationale. To repeat, in the reflexive case I do not "resolve to heap reproach on myself if my act is a violation of duty[...] to precipitate the causation of a better act than would occur otherwise".[50] First, heaping reproach, feeling guilt, etc. are not a matter of decision. But even more importantly, second, what happens

[48] Ibid., 240.

[49] Johnson, 'The Authority of the Moral Agent,' 269.

[50] Arneson, 'The Smart Theory of Moral Responsibility,' 241.

is that my "heaping reproach on myself" induces me to act better next time, if it does, because I reproach myself for non-instrumental, non-strategic reasons just because what I did was wrong. The same point applies to holding others responsible. Holding others responsible achieves its purpose, if it does, because holding others responsible is predicated on judgements of responsibility.

The reasons for "blaming [others] for misdeeds and praising them in a judging style"[51] are therefore not "pragmatic". If that were the case, these manifest responses to agents would not work at all, except when they involve harsh treatment of the agent but then it is the treatment that exerts all the influence and not the judgement. If consequentialists are prepared to go that far, however, then it might be the theoretically more honest option to do away with the notion of responsibility altogether.

In light of these objections, I conclude that consequentialism is either self-defeating or skeptical about responsibility. Self-defeating if it holds on to any notion of responsibility as a judgement *of* the agent because the very forward-looking concern it advocates cannot be promoted other than by a judgement the appropriateness of which is not dependent on the forward-looking concern. Or skeptical because it cannot accommodate in its theoretical scheme ascriptions of responsibility as judgements *of* the agent.

3.5 The truth in consequentialism

Despite these criticisms, I would like to argue that the consequentialist theory of moral responsibility contains a number of insights that, although not quite in the way understood by consequentialists, capture important aspects of responsibility-attributions. We can come closer to seeing what those insights are by appreciating the consequentialists' insistence that the notion of responsibility acquires meaning only in a social context. Instead of speaking about social context it may be better to refer to the "interpersonal embeddedness" of imputations of responsibility since it is not a given society or concrete social arrangement that one has in mind. Nor should it be implied that the validity of responsibility-ascriptions is dependent on social conventions or on their being actually practised in any given society. What is meant is simply that the true significance of being responsible can only be captured against the backdrop of agents interacting with one another. If we are to understand the concept of moral responsibility we have to ask why we *need* that concept at the first place and it will be found that we need that concept because we interact with one another. All its considerable shortcomings notwithstanding, the consequentialist account is predicated upon that understanding and that understanding is, I believe, worth preserving for reasons to be discussed in this section.

[51] Ibid., 240.

The basic difficulty with the consequentialist account, on the other hand, is its understanding of *what* we need that concept for. Thus I have argued that the consequentialist theories discussed above run into difficulties because they attempt to construe the *point* of particular responsibility-attributions exclusively in terms of the forward-looking concern. The picture painted of individual ascriptions of responsibility as instruments of social control was found to be implausible.

But perhaps there is also a more sophisticated consequentialist approach that does not seek to justify every particular ascription of responsibility by appeal to a consequentialist calculus while still justifying the general practice of responsibility-attributions in terms of the forward-looking concern? The advantage of this approach would lie in its ability to provide a plausible consequentialist account of the rationale of the general practice of responsibility-attributions in terms of the social function of that practice while avoiding the mistake of trying to derive the justification for particular ascriptions of responsibility directly from this function.

Such a two-tiered consequentialist approach is spelled out, among others, in an early work of John Rawls.[52] His account is based on distinguishing "between justifying a practice and justifying a particular action falling under it".[53] The point is that while the practice itself may be justified on consequentialist grounds, justification in particular cases need not invoke consequentialist considerations. For example, there is no inconsistency in saying that the institutional practice of punishment is justified *on the whole* because it deters wrongdoers but punishment of wrongdoing in any given particular case is retributive, i.e. the criminal *deserves* to be punished because he broke the law.[54] By the same token, the practice of promising is justified on consequentialist grounds, but this does not mean that consequentialist arguments will figure in an explanation of why any given promise is to be kept. Nor will I be allowed to invoke such arguments to justify breaking a promise.[55]

It is important to note that this consequentialist approach is very different from the kind of "Government House utilitarianism" defended by Arneson in Sidgwick's footsteps. Rawls's idea here is *not* that it is in society's interest that the masses of those engaged in everyday moral practices be left in ignorance of the true justification of their judgements. Rather, Rawls allows that justification in particular cases is purely deontological. There is no hidden agenda. The reason why I ought not to break a promise is that it amounts to a violation of a duty, not that doing so threatens to undermine the institution of promising. The ground for punishing a criminal is that he is guilty, not that punishing him may produce beneficial consequences.

[52] Rawls, 'Two Concepts of Rules.'

[53] Ibid., 20.

[54] Ibid., 22.

[55] Ibid., 29-33.

At the same time, Rawls says that the general practice is to be justified on consequentialist grounds. But will such a general justification not commit the consequentialist to accepting punishment of the innocent? Rawls argues that this unwelcome implication does not follow. The consequentialist can show that in the legal arena the forward-looking concern is best promoted by setting up the institution of punishment in a way that only those guilty are punished. If officials were authorized to condemn the innocent whenever they deem that to be in society's best interests, the results would undermine the institution of punishment itself. The collapse of the institution is likely to happen for at least two reasons. First, because under such an arrangement officials entrusted with administering punishment could easily come to pervert the institution in pursuit of their own interests. And second, because people could be expected to develop ambiguous attitudes towards the institution and fail to cooperate with it. As a result, the institution could no longer serve its purpose and work against rather than promote the interests of society. The upshot is that consequentialist arguments themselves favour setting up the institution of punishment in a way that only the guilty are punished and further in a way that officials are continuously monitored, rules are made public, authorization rests on formal procedures, etc.[56]

This two-tiered approach is not limited to practices regulated by formalized institutions as in the case of legal punishment. It can be applied in the moral domain as well. Thus there are sound consequentialist arguments why "the point of having the practice would be lost"[57] if the practice did allow consequentialist arguments as an excuse for breaking a promise. In the same vein (although Rawls does not do so), one could apply the legal analogy to the practice of ascribing moral responsibility as well. One could make the case that if the practice were such that it did not track the agent's blameworthiness or praiseworthiness the very purpose of the practice would be undermined. It would follow that particular judgements of moral responsibility are justified if and only if the agent is blameworthy or praiseworthy. At the same time, the practice of responsibility-attributions on the whole would still be justified on the grounds that the practice serves society's interests by deterring wrongdoers and encouraging those who observe their duties.

There are a number of powerful and familiar criticisms of this approach. Meanwhile, I also think that it can help to save the above mentioned insight concerning the embeddedness of responsibility-attributions in social contexts. But for that the unnecessary consequentialist baggage must be first shed. Therefore, I will begin by reviewing the shortcomings of the two-tiered approach.

[56] See ibid., 27-8.

[57] Ibid., 32.

Rawls suggests that we read the distinction between justifying a practice and justifying a particular action as a *logical* one. According to this suggestion, there are "two concepts of rules". The summary view pictures rules as heuristic tools–"rules of thumb"–helping to save time or guide action in recurrent similar cases. But there is another concept too, namely the practice conception, upon "which rules are pictured as defining a practice".[58] This conception applies to actions which are only made possible by the relevant rules being in place. For example, only once the rules of chess are in place will it be possible for you to perform the action known as 'checkmating your opponent'.

The most important difference between the two conceptions of rules[59] is a difference in their justificatory potential. On the summary conception, the appeal to the existence of a rule does not have independent justificatory significance. The rule sums up how past cases have been decided but gives no reason in itself to decide the case in one way or another. The rule does not weigh in the balance of reasons. By contrast, rules under the practice conception can very often themselves decide what is right to do in the case at hand. When one's actions of this type are challenged, "one doesn't so much justify one's particular action [e.g. keeping a promise, imposing punishment] as explain, or show, that it is in accordance with the practice."[60] Pointing to the rule in this way can be sufficient to decide the matter.

It is the practice conception that adequately describes, so Rawls, practices such as that of promising or punishing. When one is asked to justify keeping a promise or imposing punishment on someone, one merely needs to point out that it is in accordance with the rules of the relevant practice. But, as we have seen, these are such that, barring emergencies, promises are to be kept independently from how the sums of the consequentialist calculus come out and punishment is to be imposed only if the agent is guilty-as-charged, again independently from whether the sums will come out right. Only if the practice is questioned as a whole will it be necessary to review the consequentialist arguments in its favour.

[58] Ibid., 36.

[59] In the following I am going to use terms 'rule' and 'norm' interchangeably. Raz points out that not all rules are norms (see Raz, *Practical Reason and Norms,* 9, 117) as not all rules have the action-guiding function attributed to norms. For example, the rule specifying the number of players on a football team does not directly generate reasons for action. Whenever the term 'rule' is used in the following it is meant to refer to normative rules, however. Note also that, coupled with their action-guiding function, norms or normative rules also have an evaluative function in the sense pointed to by Wiggins when recalling the etymological origin of the word (see Wiggins, *Ethics,* 236): "The word *norma* meant for the Romans a T-square that a carpenter or mason carried about with him for making right angles". Thus a norm or normative rule is a *fixed* standard against which a particular instance falling under the norm is measured or assessed.

[60] Rawls, 'Two Concepts of Rules,' 36.

The first major criticism of this account is that it is a mistake to defend the two-tiered approach on logical grounds. For one thing, as noted by Joseph Raz, there is a general problem with the feasibility of the distinction itself between the two concepts of rules.[61] The problem is that so long as a rule functions as a rule, i.e. it is perceived to be valid and pertinent to the given situation, the proposed distinction between the summary conception and the practice conception will not be relevant to how rules generate reasons. It is true that, to take Rawls's example, "someone's being fatally ill and asking what his illness is, and someone's telling him" are all things that can be described regardless of whether or not there is a rule to the effect that one should not inform a patient about his illness being fatal.[62] But the point is that if one believes that there is such a rule, then it will necessarily enter into the balance of reasons. One may (justifiably or not) decide to override that rule in the end, but the rule is not a rule if the belief that there is such a rule does not generate reasons.

The situation is not relevantly different in the case of rules which Rawls thinks fall under the practice conception. The belief that there is a rule specifying that a given position amounts to checkmate will provide reasons. But of course, just as in Rawls's example of informing an ill patient, we can describe any position on the chessboard by specifying the location of the pieces without saying that that position is checkmate according to the rules of chess. We can even prepare a complete list of positions on the board which are such that the opponent's king will be taken. What cannot be done is to explain the players' actions without invoking the rule that that position is checkmate, i.e. ends the game. This is because the belief that there is a certain rule of checkmating generates reasons.

But, again, precisely same holds for not informing the patient in Rawls's example. One may of course decide not to inform a terminally ill patient about his condition for a variety of reasons other than one's belief that there is a rule to that effect (e.g. one may just not feel strong enough to break the horrible news). But as long as the information is withheld *because* one believes that there is such a rule, the rule will function in the same way as the rule of chess in the sense that it will enter into the balance of reasons by virtue of being a rule believed to be pertinent to the situation. Of course there may be many other kinds of differences among types of rules: in terms of scope, function, relation to other rules and, most importantly, in terms of *why* they are regarded as pertinent to the situation. I am only denying here that the distinction between the practice and summary conception is relevant to the problem of how rules generate justificatory reasons. In sum, we can agree with Raz's conclusion that while to fully explain the normative significance of some actions we "must include reference to a rule[. . .] this

[61] Raz, *Practical Reason and Norms,* 108-11.

[62] Rawls, 'Two Concepts of Rules,' 35.

distinction between acts or their descriptions does not lead to a corresponding distinction between types of rules".[63]

In any case, even if it were possible to draw a general distinction between types of rules on those logical grounds, it is questionable whether that distinction could be used for the specific purpose that Rawls had in mind. The worry here is succinctly put by Conrad D. Johnson: "Not being free to save innocent lives by killing an innocent is not a stricture imposed on us by logic."[64] In other words, what is wrong with the defence that punishing an innocent person or breaking a promise seemed best in view of the expected consequences is not that such a defence entails a category mistake. In fact so much is admitted by Rawls himself in the passages where he argues that breaking a promise is, ceteris paribus, inherently wrong ("The promisor is bound because he promised: weighing the case on its merits is not open to him"[65]) or where he requires guilt to justify punishment.[66]

This does not mean that the distinction between justifying a practice and justifying an action falling under it must be abandoned. But it does mean that in the cases discussed by Rawls it will not be possible to argue that a certain course of action is unjustifiable because acting in that way entails misunderstanding the practice or misunderstanding what it is to engage in that practice. Breaking a promise or punishing the innocent for the reason that it seems best to do so from a consequentialist perspective is not conceptually or logically but morally objectionable (if at all). If you promise to pay your debt to me by the end of the month and you do not, I will not question your understanding as I would if you had made an illegal move in chess but rather (in the absence of a legitimate excuse or defence) I would begin to doubt your moral mettle.

It is my impression that Rawls himself is aware of this problem which can be seen from the fact that he gestures towards two quite different arguments to defend the two-tiered approach to justification. *First,* a number of passages suggest that the two-tiered approach is to be read as a plea for radically extending and refining our understanding of how best to promote the Utilitarian good. This is most notable in his discussion, already discussed, of why the institution of punishment should be set up in a way that avoids punishing the innocent. As has been seen, he rejects the institution of "telishment" ("which is such that the officials set up by it have authority to arrange a trial for the condemnation of an innocent man whenever they are of the opinion that doing so would be in the best interests of society")

[63] Raz, *Practical Reason and Norms,* 110.

[64] Johnson, 'The Authority of the Moral Agent,' 272.

[65] Rawls, 'Two Concepts of Rules,' 32.

[66] Ibid., 22.

on the grounds that a "utilitarian justification for this institution is most unlikely".[67]

The difficulties with this defence of the two-tiered approach are fairly obvious. Non-consequentialists would reject "telishment" for quite different reasons, namely because they think that is inherently wrong to punish an innocent person. Even if there were a consequentialist justification of punishing an innocent person we ought not to do so according to non-consequentialists. Moreover, it cannot and should not be a mere coincidence that consequentialist and non-consequentialist intuitions coincide, if they do, regarding how we are to set up the institution of punishment. As I have indicated above, Rawls himself accepts that (barring emergencies) punishing the innocent or breaking a promise are inherently wrong in other parts of his essay. His problem is precisely to reconcile these deontological intuitions with a consequentialist justification of the general practice. Extending and refining the analysis of potential consequences will not be sufficient to meet this challenge.

In addition, although it is true that there are consequentialist reasons too for not punishing the innocent, it is very unlikely that from a consequentialist perspective the institution of punishment would always work best if and only if the guilty are punished. But if that is correct, then it is hard to see how Rawls's proposal can help in fending off the problem at the heart of rule-consequentialism. Either following the rule ('punish only the guilty') will lead to sub-optimal consequences. Or if it does not, then that will be because what the rule tells the agent to do coincides with what the agent was to do anyway if he were to act optimifically in the given situation. In any case, the rule seems to do no independent work towards furthering the forward-looking concern and therefore the question of justification must be limited to particular actions only. But from that it would naturally follow that the distinction between justifying practice and justifying a particular action falling under it collapses too. The same point applies to promising or any other formal or informal practice. Therefore, *this* argument will certainly not give us the best of both worlds, i.e. simultaneously justify the overall practice on consequentialist grounds while rescuing intuitions about the pertinence of non-consequentialist considerations to particular cases.

There is, however, a *second,* more promising line to argue for the two-tiered approach hinted at in some passages by Rawls and spelled out in more detail by Conrad D. Johnson.[68] According to this argument, the two-tiered approach is best read as calling for a moral division of labour. On this reading, the distinction between justifying practice and justifying a particular action falling under it is based on a distinction between different offices or roles. Justifying the practice is a task that falls only to a certain

[67] Rawls, 'Two Concepts of Rules,' 27-8.

[68] Johnson, 'The Authority of the Moral Agent.'

kind of office or role, but not to other kinds which only involve justifying particular cases.

The legal analogy, which Rawls himself also relies on, is quite helpful here. The office of the judge is to decide disputed cases in accordance with pre-existing law. Only the legislator, who holds a different office, is authorized to change the law itself. The legislator may decide to do so to promote various goals. But those goals are beyond the judge's purview insofar as the judge's decision is to follow existing legal prescriptions even if *not* following them would be more conducive to the general goals of the legislator.[69]

In other words, what office or role he assumes will limit the agent's competence as to *what kind of reasons* he can legitimately invoke to justify his action. For instance, the judge does not have the authority to justify a decision contrary to pre-existing law by claiming that the decision seems more in line with the legislator's original intent. Authorization, therefore, is about the kind of reasons one can legitimately have access to. But note that authorization attaches to the office and as such it is independent from the content of the legal norms that the judge is to consider in reaching his decision. That is to say, the judge is not authorized to ignore existing legislation when evaluating the merits of the case. Consequently, certain considerations will be automatically excluded from the balance of reasons on the basis of which he will adjudicate any given case.

It complicates matters but is not an objection to this proposal that judges are sometimes required to create new law or innovate existing legislation. Raz argues that this happens when the "law is unsettled", i.e. when it is unclear what judicial decision existing legal sources require.[70] If Raz's analysis is correct, then there arises no real difficulty for the division-of-labour interpretation of the two-tiered approach since judges only assume the role of the legislator, or something akin to it, precisely when there is no pre-existing norm applicable to the particular case or it is particularly difficult to determine what the norm requires or when existing norms conflict. But according to the argument above the division-of-labour becomes effective only if there are norms in place regarded as valid and reasonably unequivocal by the participants of the practice to which these norms apply.

This second way of defending the two-tiered approach assumes that a similar division of labour exists between the role of the moral agent and that of the moral legislator.[71] Because the moral agent stands under the authority of the putative moral legislator, there are certain things the moral agent cannot do. Thus he is not free to revise the "moral code",[72] i.e. the set of rules in terms of which action is morally assessed. That is not to say that anyone is in principle barred from assuming the role of the moral

[69] Johnson, 'The Authority of the Moral Agent,' 272.

[70] See Raz, 'Legal Positivism and the Sources of Law,' 49-50.

[71] See esp. Johnson, 'The Authority of the Moral Agent,' 278-9.

[72] Ibid., 271.

legislator. But when in the role of the moral agent, one cannot challenge the normative framework itself that limits the kinds of reasons the agent can have access to in order to justify his action. That framework itself is imposed by the authority of the moral legislator.

The term 'moral legislator' is of course not meant to designate a single person or decision-making body. The label can stand for any kind of authority–including the impersonal authority of convention or that of shared values–from which the norms emanate, the norms that is, relative to which justification in particular cases proceeds. In fact, the division of labour proposed here is best conceived of as a distinction between positions, the position of the agent/judge and the position of the legislator/deliberator who can assess and contest the norms themselves. The autonomy of participants in the practice remains intact as long as they have equal access to both the position of the moral agent and that of the legislator. The crucial point is, however, that "the moral legislator and the moral agent are *literally* not the same entity in the same place and time".[73] If read correctly, this is a point about how normative authority, any kind of authority, operates. How such authority is to be justified is a separate question (of which more below).

In addition to its general appeal, this division-of-labour conception offers the prospect of reconciling the consequentialist justification of a general practice with the justification of deontological intuitions applying to particular cases falling under that practice. Thus, for example, when confronted with the question whether it is justifiable to break a promise, it is beyond the agent's purview to undertake a full consequentialist evaluation of the case. This is not because to do so would entail misunderstanding the practice of promising, but rather because *qua* moral agent he is not authorized to do so. Similarly, when confronted with the question whether to impose punishment on someone or not, it is beyond the (moral) judge's purview to consider how the consequentialist calculus would come out if an innocent person were to be punished. Once again, this does not mean that anybody would be in principle barred from undertaking the consequentialist assessment, but "*this* agent in *this* situation must adhere to restriction R without allowing consequentialist considerations any full impact... *That* question, though open on other occasions, is closed for this agent here and now".[74]

I believe that not even the division-of-labour interpretation can deliver the hoped-for reconciliation of the consequentialist justification of a practice with the deontological intuitions applying to particular cases falling under that practice. At the same time, the most basic insight of the two-tiered approach, i.e. the distinction itself between justifying a practice and justifying a particular action falling under it, can be preserved. The division-of-labour

[73] See ibid., 279.

[74] Ibid., 275, 278.

interpretation goes a long way in showing how this can be done. This is because it recognizes that depending on one's position (that of the agent as opposed to that of the legislator) one has access to a different order of reasons for the purposes of justification. Even if true, however, this finding is of general interest for moral philosophy and therefore only indirectly pertinent to the cognitivist conception of responsibility at the center of attention here, so I will not take it up in the following chapters.

The division-of-labour reading of the two-tiered approach does not in itself produce arguments in favour of a consequentialist grounding of the practice of responsibility-attributions. Even if successful it *could* only show that a consequentialist account of the general practice is in principle compatible with a non-consequentialist justification of particular ascriptions of responsibility. In other words, it *could* show that there need not be an inconsistency in saying that the general practice of responsibility-attributions serves to promote the forward-looking goals of deterrence and encouragement, while the justification of ascriptions in specific cases remains independent of those goals.

But the division-of-labour conception produces very little by way of positive argument in favour of the claim that a consequentialist rationale would underlie the practice of responsibility-attributions as their normative foundation. In fact, the same considerations militate against embracing such a rationale as those which have been mentioned in connection with other versions of rule-consequentialism. Analogously to the two-tiered analysis of punishment, the two-tiered understanding of responsibility would call on the moral legislator to shape the practice of responsibility-attributions in a way that only those who *are* in fact responsible will be *held* responsible. If someone is not in fact responsible it will be impermissible to hold him responsible (barring emergencies), even if that would promote the forward-looking concern.

The problem is, however, that if the priority of being responsible over holding responsible is preserved in particular cases, then it is hard to see how that order could be reversed in justifying the general practice. As already noted, it is implausible to claim that the general practice of responsibility-attributions would promote the goals of deterrence and encouragement most effectively if in particular instances only those are held responsible who are in fact responsible. As a result, what would be justifiable according to the optimific rules of the practice will clash again and again with what seems justified in particular cases. This will happen not simply "because the moral legislator may have made a mistake, or the moral code may not have been revised recently enough",[75] but because the considerations invoked for the purposes of justification are at loggerheads. However, if no such reconciliation of the consequentialist justification of the overall practice and the

[75] Ibid., 279.

non-consequentialist justification of particular cases falling under the practice is possible, then we either have to apply consequentialist justification directly to particular cases as well or give up the attempt to provide consequentialist justification for attributions of responsibility at any level. The former option has been already rejected for reasons discussed in Section 3.4. So there seems to be no alternative left but to dismiss consequentialist justification of responsibility-ascriptions both for ascriptions in particular cases and for the practice of attributing responsibility as a whole.

At the same time, the division-of-labour reading of the two-tiered approach can be used to defend the distinction between different levels of justification. The merit of this approach is that it makes it clear that the authority of a valid rule or norm for the agent consists not only in its ability to tip the balance of reasons but also in limiting the range of reasons to be considered. The agent who is to act in a particular case, or the judge who is to judge a particular case, is in a normatively different position from the legislator or the critic who is to evaluate the normative framework. Once in the position of the judge or the agent, we are not free to alter the normative framework. As long as the autonomy of individuals is uncurtailed, they must have access to both positions in equal degree. But they cannot occupy both positions simultaneously. If they are in the position of the agent/judge, the normative framework constituted by a rule or a set of rules will have authority for them. They may contest that authority at any given time but by doing so they will assume the legislator's or the critic's role and move away from that of the agent/judge.[76]

I believe that this general conceptual scheme applies to the problem of responsibility as well, but I will not explore this avenue any further in this work. In any case, it is important to emphasize that the above considerations do not concern the justification of the authority of rules or norms but rather are about what that authority consists in. There can be a fairly wide variety of reasons for adopting rules or norms and so there can be various justifications for regarding them as valid. Following a rule may be thought to produce the best consequences overall or following the rule may be thought to be one's duty and there may be other reasons too. So too, even though the division-of-labour reading of the two-tiered approach was originally proposed to reconcile a consequentialist justification of rules constituting a practice with deontological intuitions applying to particular cases,

[76] Consider also this example (such analogies are made possible by the fact that the difference in justificatory positions exists in other normative domains as well): There is a difference between the position of a chess player debating whether a certain arrangement of the pieces on the board is indeed checkmate or not, and the position of officials of a chess association debating whether the rules of chess should be altered so as to extend the number of arrangements which are to be counted as checkmate (for example, one could argue that it would do better justice to the spirit of the game to re-classify stalemate as checkmate with the victory of one player rather than a tie).

the point about different justificatory positions is quite independent from this particular application. In other words, the claim about the division-of labour can stand even if the justification for having rules is not conceived of in consequentialist terms. Further, it was also noted that this point is not tied to a logical distinction between types of rules (as originally thought of by Rawls), but rather is meant to describe how rules generate reasons.

3.6 Conclusion

In this chapter I have first rejected attempts to justify responsibility-ascriptions on directly consequentialist grounds (act-consequentialism about responsibility) and then I have also criticized theories which seek to account for the norms or rules under which such ascriptions fall in consequentialist terms (rule-consequentialism about responsibility). At the same time, the discussion was intended to expose two points elaborated by at least some versions of the consequentialist theory which, I suggest, can be adopted by a cognitivist understanding of responsibility otherwise unsympathetic to the consequentialist outlook.

One such insight, the *first* truth in consequentialism, was the 'interpersonal embeddedness' of responsibility-ascriptions. Consequentialism seems to capture quite effectively an essential truth about our notion of moral responsibility being tied up with our participation in practices which involve transactions with other agents (and derivatively with oneself too). Consequentialists talk persuasively about the "good of the blaming" (Arneson) and generally about the good to be had from engaging in the practice of responsibility-attributions. I believe that consequentialism has an important point to make by insisting on the value of responsibility and by relating this value to interactions between persons.

I rejected the consequentialist priority of 'holding responsible' over 'being responsible' for both descriptive and normative reasons. But the fact that judgements of responsibility enjoy both psychological and justificatory priority over actual, overt responses to agents (the Priority Thesis) should not be taken to mean that these judgements are driven by a merely theoretical interest. Rather, we pass such judgements all the time because we interact with one another and we would like these interactions to take a certain shape even if they do not affect us directly. In this sense it is not meaningless to talk of the *point* or *good* of imputing responsibility in human interactions.

It is quite another matter that consequentialism appears wrongheaded in its characterization of that point or good, i.e. the rationale or concern underlying responsibility-ascriptions. Thus we have found good reasons not only to question the relevance of consequentialist considerations to the justifiability of responsibility-ascriptions in particular cases, but also to question the more sophisticated consequentialist claim that while in particular cases the ascription of responsibility need not be consequentialist, the overall justification of the practice of imputing responsibility could nevertheless be based on a forward-looking concern. No such reconciliation of deontological intuitions and consequentialism seemed possible.

But if I am right that the consequentialist characterization of the point of the practice of responsibility-attributions is mistaken, it will be necessary, beyond the criticisms made in this chapter, to offer an alternative characterization of that point. If it is true that that point is quite independent from the concern for the maximization of expected utility, what good, if any, is to be had from regarding one another and ourselves as responsible agents? In Chapter 6 I will try to answer that question by arguing that responsibility is something we value as an essential aspect of personhood.

It will also be recalled that reconciliatory versions of rule-consequentialism were found to contain another insight too, the *second* truth in consequentialism, namely the general possibility of distinguishing between levels of justification. Although it seems right to get rid of the consequentialist baggage, we can nevertheless hold on to the general distinction between justifying a practice and justifying a particular action falling under it.

In the next chapter, I will discuss Peter Strawson's theory of responsibility which also makes much of both the 'interpersonal embeddedness' of responsibility-attributions and the distinction between justification at the level of practice and at the level of particular ascriptions of responsibility. As will be seen, however, the Strawsonian conception importantly diverges from consequentialism both as regards the implications of 'interpersonal embeddedness' of responsibility-attributions and in its understanding of how these attributions can be justified.

Chapter 4

Is Responsibility Inescapable? Peter Strawson's Naturalist View of Moral Responsibility

4.1 Introduction

One may perhaps wonder whether it is not disproportionate to devote an entire chapter to an account of moral responsibility which is spelled out in the breadth of a single, fairly concise essay, Peter Strawson's *Freedom and Resentment.*[1] But this work, quite apart from its sheer elegance and zest, is so rich in ideas and has inspired so much posterior reflection that it can rightly be called a new beginning in how philosophers think about responsibility.

The most important novelty of the Strawsonian approach lies in the connection it establishes between the concept of moral responsibility and the world of human emotions. The exact nature of this connection has later been construed in different ways, but no theory of moral responsibility after Strawson could ignore it. Strawson's most important contribution is to have shown that attributions of responsibility typically involve more than affectless pronouncements of blame and praise. When one person blames another his response is, more often than not, dominated by negative feelings of resentment, anger, indignation, scorn, etc. Likewise, when one person praises another his response is characteristically marked by gratitude, admiration and satisfaction. Because these emotions are principally triggered by

[1] Strawson, 'Freedom and Resentment.'

what someone else has done (or is perceived to have done), they are rightly referred to as *reactive* sentiments.[2]

What Strawson is at pains to emphasize is that the connection between reactive sentiments and the notion of responsibility is not a contingent feature of human psychology. To understand the concept of responsibility, we must understand the nature of these reactive sentiments and the attitudes they express. Not only will this throw light on why we care at all about responsibility and agency, but will also allow us to make headway on the conundrum at the heart of the freewill debate, namely the relevance of determinism to responsibility. What's more, Strawson expects to recast our entire conception of responsibility by proposing to think of responsibility-attributions as affective phenomena: "Only by attending to this range of attitudes can we recover from the facts as we know them a sense of what we mean, i.e. of *all* we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice."[3]

It will be recalled that for the purposes of grouping alternative theories of moral responsibility I have relied on two basic distinctions: (i) holding vs being responsible, and (ii) emotional response vs cognitive content. Attempting to position Strawson's account in the resulting 2x2 classificatory matrix can again prove to be a helpful point of departure. It might seem that Strawson's insistence on understanding responsibility in terms of affective *reactions* to agents places his theory firmly among those which define responsibility on the basis of behavioral dispositions, that is, explicate attributions of responsibility as something people do in response to what is being done to them. It might also seem that his insistence on understanding responsibility in terms of *affective* reactions, aligns the theory with those which explicate attributions of responsibility primarily in terms of the emotions (e.g. guilt) typically accompanying such attributions rather than in terms of the attributions' propositional content.

As will be seen, matters are not that simple. First, although no one can deny that establishing the connection between responsibility and reactive sentiments is a key achievement of the Strawsonian account, commentators have differed on the best reading of the conclusions Strawson intends to draw from this connection. This difference of opinion is understandable as, second, in *Freedom and Resentment* Strawson makes several different suggestions (all intriguing but not all mutually compatible) as to how to interpret the role of reactive emotions. So the first task I would like to tackle in this chapter is to reconstruct Strawson's main arguments and thereby position 'Strawsonianism' relative to alternative theories of moral responsibility (Section 4.2). I will also spend some time on Strawson's characterization of

[2] For my construal of reactive sentiments as normative consequences of responsibility-ascriptions, see Chapter 2, p. 35f.

[3] Ibid., 23.

reactive attitudes and their supposed opposition to what Strawson refers to as the "objective attitude" (Section 4.3). Then I will move onto the question how dominant a role emotions play in shaping these reactive attitudes (Section 4.4). The same section will address the problem of accounting for the special normative significance of certain reactive attitudes (referred to as "moral reactive attitudes") and whether such an account can still be accommodated within the framework of a Strawsonian theory. My answer will be in the negative.

Of course, as evidenced by the citation two paragraphs further above, Strawson does not rest content with pointing out the connection between responsibility and moral sentiments. Quite the contrary, he makes a number of far-reaching claims about responsibility of both metaphysical and normative import which, Strawson thinks, depend crucially on this connection not being a contingent one. In particular, *Freedom and Resentment* formulates a highly influential view concerning the justification of responsibility-attributions. This view is supported by a number of different and to some extent contradictory arguments all of which focus on the notion of *inescapability.* These arguments will be critically discussed in Section 4.6. It is also in this section that we can return to the problem already broached in connection with consequentialism,[4] namely what it means for the explication and justification of responsibility-attributions to be embedded in a practice. Strawson has valuable contributions to make to the discussion of this problem as well.

One issue that I promised *not* to make the central concern of this work is that of the compatibility of the truth of determinism with moral responsibility. Since, however, Strawson's position is highly innovative on this issue as well (quite different, for example, from the consequentialist brand of compatibilism briefly introduced in the previous chapter[5]) and because what he has to say about this issue could only be artificially disentangled from what he says about the other issues listed above, I will also engage with his brand of compatibilism as well (Sections 4.2 and 4.5).

4.2 Strawson's reconciliatory compatibilism

In this section, I would like to review the argumentative backbone of the Strawsonian conception and sum up Strawson's main conclusions. This can help to map out the terrain which I would like to explore in detail throughout the following sections.

Step one. Strawson begins by making a "commonplace" observation about the normal course of interpersonal transactions: participation in interpersonal human relationships is marked by a distinct range of reactive

[4] See Chapter 3, Section 3.5.

[5] See Chapter 3 on the Compatibilist Thesis, p. 45.

feelings and attitudes. These feelings and attitudes are prompted by our perception of the degree of goodwill or regard manifest in other agents' actions. Thus we feel and display resentment, indignation, anger, gratitude, satisfaction, and more remotely, love or forgiveness depending on what a given action reveals about the agent's intention or attitude towards us. We do sincerely care about the agent's attitude and not only the benefits or injuries caused by the action itself. This is shown by the fact that we often entertain the said reactive attitudes even in response to, say, the otherwise harmless "manifestation of attitude itself" by a rude or indifferent person.[6]

Step two. Such participatory reactive attitudes are contrasted with the objective attitude. The objective attitude is analyzed by Strawson in terms of, first, the considerations which move one to adopt it and, second, the different emotional/behavioural patterns displayed by those who adopt it (as will be seen later on the proposed terms of the analysis are to some extent contradictory). Thus, first, in order to understand the objective attitude we can begin by looking at the certain special considerations which induce us to take it. These considerations typically involve an assessment of the agent's capacities as being "abnormal" or "undeveloped" or an assessment of the given action as not reflecting the agent's true attitude towards us because the agent couldn't help doing what he did or did not know what he was doing.[7] On the other hand, second, we can also attempt to characterize the objective attitude by considering how our patterns of behaviour and emotional dispositions change when we assume this attitude. According to Strawson, what is common to the patterns and dispositions associated with the objective attitude is the lack of participatory reactive sentiments. From the perspective of the objective stance the other agent can be seen as an "object of social policy" or as a subject of treatment, "something to be managed or handled or cured or trained; perhaps simply to be avoided". But the objective attitude will never involve resentment, gratitude, forgiveness, anger, etc.[8]

Step three. The objective attitude is not something we can assume as a *universal* stance towards others (or ourselves). Most commonly, we assume it in response to severe psychological abnormality. We might in some instances take the objective attitude towards 'normal people' as well, but in any case, the withdrawal from interpersonal relationships into the objective attitude can only be temporary and cannot last very long.

Step four. Whatever our position on the thesis of determinism we will never be prompted by this thesis to take the objective attitude. The reason

[6] Ibid., 5-6.

[7] Ibid., 8. Following Gary Watson, I have previously referred to global responsibility-undermining conditions as 'exemptions' to distinguish them from local 'excuses' such as ignorance or compulsion. See Chapter 2, p. 13. and Watson, 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,' 259-61.

[8] Strawson, 'Freedom and Resentment,' 9.

is the following. The considerations which do prompt us to assume an objective stance are special in that they apply when either the agent or the circumstances of the action diverge significantly from the norm. If we did accept that determinism itself could be a valid responsibility-undermining condition, then in effect we would be *generalizing* special considerations as applying to all actions of all agents. But by definition there can be no such thing as a universal excuse/exemption[9] because excuses/exemptions are valid precisely because something is atypical or abnormal about the agent or the circumstances of the action. If determinism was relevant to what attitude we go in for, this would entail that we were to see "abnormality as a universal condition". That, however, is an evident self-contradiction: abnormality cannot be the universal norm. Therefore determinism cannot be relevant to how we react or how we should react to others.[10] Q.E.D.

Step five. Q.E.D.? Not quite or at least not yet. Strawson admits the possibility that taking refuge "from the strains of commitment" or "as an aid to policy" or "simply out of intellectual curiosity" we may temporarily opt for the objective attitude vis-à-vis normal, mature persons as well. Granted, he argues that taking the detached, objective view of the behaviour of "the normal" and "the mature" typically only serves to review the terms of engagement for the purposes of reconciliation or explanation. If this cannot be carried out, then we sever the relationship rather than persist in the objective attitude.

That may be true, but this is nevertheless a crucial concession because it admits reasons for disengaging from the ordinary range of reactive attitudes even vis-à-vis agents who are not psychologically incapacitated or underdeveloped. What Strawson needs to do at this point, therefore, is to show that the truth of determinism does not constitute such a reason. It will not do to appeal to a putative self-contradiction because by Strawson's own admittance now there is no self-contradiction involved in relating to "the normal" and "the mature" as if they were incapacitated.

Step six. So we should not be surprised to find that Strawson advances (at least) three distinct arguments beyond the *argument from self-contradiction* (made in Step 4 above) why determinism cannot be a reason to assume the objective attitude or in fact a reason to alter our attitudes in any way towards others. Although they are presented conjointly and are evidently supposed to reinforce one another, strictly speaking each of these arguments is independent from one another in the sense that each one is deemed by Strawson to be sufficiently strong on its own to establish the irrelevance of determinism. These arguments are the following:

[9] For the distinction between excuses/exemptions, see Chapter 2, p. 13 and note 7 above.

[10] See Strawson, 'Freedom and Resentment,' 11.

- *The argument from naturalism:* giving up our commitment to reactive attitudes and feelings would entail withdrawing from the world of interpersonal relationships. But that is practically inconceivable because the commitment to reactive attitudes and feelings is a 'given' of human nature. This is also why we are not able to assume the objective attitude as a universal stance (see Step 3 above). If that is true, however, it is simply irrelevant whether or not the truth of determinism would require us to abandon reactive attitudes and feelings.

- *The argument from value:* abandoning reactive attitudes and feelings, which are part and parcel of interpersonal relationships, would amount to repudiating something that is of *value.* We are clearly better off having these attitudes and feelings given their unique potential to express our interest in others' goodwill (or the lack of it) towards us and our goodwill (or the lack of it) towards them. In addition, morally-significant reactive attitudes also cater to a more general concern. Being the "sympathetic or vicarious or impersonal or disinterested or generalized analogues"[11] of the personal reactive attitudes, moral reactive attitudes are the irreplaceable vehicles for the expression of our interest in the general welfare of human beings. But if on the whole reactive attitudes and feelings constitute such a central value in our lives, then the truth of determinism can never constitute a strong enough reason to give up these attitudes and feelings even if we could conceivably do so.

- *The argument from rationality:* even if we were *per impossibile* constitutionally capable of giving up reactive attitudes and feelings and even if this choice did not leave us without a fundamental value, it would not be rational for us to abandon reactive attitudes and feelings. One misconstrues the nature of human rationality if one thinks that the "truth or falsity of the general theoretical doctrine" of determinism can bear upon the justifiability of responsibility-ascriptions and that of the concomitant reactive attitudes and feelings. Whatever theoretical conclusions it may produce, determinism does not yield conclusions to be taken into account in our practical choices and deliberations. The thesis of determinism leaves intact the reasons we have for our commitment to reactive attitudes and feelings because that *theoretical* doctrine cannot have any bearing on our *practical* commitments.[12]

[11] Ibid., 14.

[12] Ibid., 18-9. Note that appeals to inescapability are made by other arguments too. Thus in Section 4.6, I will also discuss the 'no justification view' according to which it is simply incoherent or wrongheaded to ask for the justification of the overall practice in which reactive attitudes and feelings are embedded. No such justification will be forthcoming because there is simply no more basic belief to appeal to beyond the bounds of this practice. What is beyond is merely the 'view from nowhere'.

Strawson concludes, therefore, that the truth of determinism remains irrelevant to our commitment to reactive attitudes. But because, on his view, reactive attitudes are constitutive of moral responsibility, it also follows that determinism is irrelevant to and therefore compatible with responsibility. To establish this conclusion has been the main purpose of *Freedom and Resentment.*

It is worth noting that this view differs from traditional versions of compatibilism subscribed to by those whom Strawson refers to as the "optimists".[13] These more familiar attempts to demonstrate the compatibility of determinism and moral responsibility have been typically based on some specific analysis of the sort of freedom agents must enjoy if they are to be held responsible (e.g. the availability of what sort of alternative courses of action, if any, is relevant to the justifiability of responsibility-attributions). Strawson's point is different. The crucial idea here is that given the inescapability of our commitments to them, responsibility-attributing practices are *immune* to the threat of determinism. Since human society is *practically* inconceivable without treating oneself and others as responsible agents, metaphysical considerations regarding the consequences of determinism are simply irrelevant to our normative *practices.*

What makes this brand of compatibilism *reconciliatory* is its pledge to accommodate both the intuitions of libertarians (to whom Strawson refers to as "pessimists") about responsibility being an essentially backward-looking and individualised notion, on the one hand, and the optimistic compatibilist idea that attributions of responsibility remain justifiable even if determinism turns out to be true, on the other.

But, as should be already obvious from the foregoing, *Freedom and Resentment* provides more than just an innovative version of compatibilism. Strawson and those inspired by his approach defend what amounts to a complete theory of moral responsibility. Or so they claim. On the basis of the above, the main theses of this theory can be summed up as follows:

1. *Constitution Thesis:* Ascriptions of responsibility are to be understood as attitudinal responses to the quality of will expressed in the agent's action. Attitudinal responses involve typical behavioural-cum-emotional dispositions referred to as reactive attitudes. Such reactive attitudes are *constitutive* of moral responsibility. That is, to hold oneself or another responsible just is to be susceptible to such reactions.

2. *Inescapability Thesis:* Although reactive attitudes can and should be temporarily suspended under certain conditions, for a number of reasons (see the three arguments above) we cannot and should not *opt*

[13] Ibid., 1.

out altogether of "the complicated web"[14] of reactive attitudes which constitute the practice of moral responsibility.

3. *Irrelevance Thesis:* Determinism is irrelevant to the practicability or justifiability of responsibility-attributions.[15]

Each of these theses relies on concepts and distinctions requiring detailed analysis which I would like to undertake in the following. I will be particularly concerned with the first and second theses, although the problems emerging in connection with these will cast serious doubt on the tenability of the third as well.

4.3 Characterizing reactive attitudes

If the above reconstruction is accurate, then the success of Strawson's proposal does seem to depend crucially on the distinction between reactive attitudes and the objective attitude. For one thing, understanding the nature of this distinction is essential to providing an overall characterization of reactiveness beyond mere examples of paradigmatic reactive attitudes such as resentment or gratitude. After all, as we have seen, Strawson claims that understanding reactiveness is the key to understanding the purpose and meaning of attributions of moral responsibility.[16] On the other hand, the distinction ought to reveal more about the objective attitude too. What view of other agents and their actions prompts this attitude? What kind of relationship to other human beings is entailed by the objective attitude? And what sort of feelings and attitudes are genuinely incompatible with objectivity?

Despite its pivotal position in Strawson's conception, at a closer look it becomes a lot more difficult to say how exactly are we to draw the reactive-objective distinction. In what follows I will review Strawson's own attempts and then consider three possible improvements on these. I will argue that none of the proposed interpretations is entirely satisfactory, but will claim that the third one contains an important clue as to the right understanding of objectivity in its opposition to reactiveness. I will conclude, however, that it is questionable whether the proposed reading of the distinction can be still made to cohere with Strawson's reconciliatory compatibilism.

As already indicated, Strawson himself draws the distinction between the reactive and objective attitudes twice over. Both of these definitions seek

[14] Ibid., 23.

[15] As will be seen, the Irrelevance Thesis also follows from the Inescapability Thesis in Strawson's presentation. But he presents at least one argument in favour of the Irrelevance Thesis, namely the *argument from self-contradiction* (see Step 4 above and p. 100), which is independent from the arguments for the Inescapability Thesis.

[16] Ibid., esp. 23.

to explain why the objective attitude (or ”range of attitudes”) is opposed to the ”range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships”.[17] My diagnosis is, however, that there is a tension between the definitions put forward by Strawson. This makes it necessary to carefully rethink, first, the distinction itself, and second, how well suited it is to support the Strawsonian brand of compatibilism.

The *first* definition Strawson provides by enumerating what count as standard *reasons* for suspending our ordinary reactive attitudes towards other agents. These reasons themselves divide into two groups. The kinds of *local* responsibility-undermining conditions previously referred to as 'excuses' belong to the first group. Such excuses affirm that the agent 'could not have done otherwise' or else was non-culpably ignorant of what he was actually doing.[18] *Global* responsibility-undermining conditions previously referred to as 'exemptions' belong to the second group. Such conditions obtain when a person is recognized as lacking a fundamental capacity necessary to qualify as a responsible agent due to being ”psychologically abnormal” or ”morally underdeveloped”.[19]

Valid excuses and exemptions constitute sufficient reasons to suspend ordinary reactive attitudes. There is a difference in the scope of suspension. In the case of excuses, ordinary responses are suspended with regard to the specific action only. Thus anger and resentment are not fitting responses to actions if the agent ”couldn't help it” or ”had to do it”. But, so Strawson, the temporary suspension of the response, does not change one's general view of the agent as one at whom we can address the ”kind of demand for goodwill or regard which is reflected in our ordinary reactive attitudes.”[20] By contrast, exemptions call for a wholesale suspension of our reactive attitudes towards the person who lacks some basic capacity necessary to qualify as a responsible agent. When we exempt a person from responsibility we take the objective attitude towards him as a general stance. This attitude is marked by the suspension of the whole range of ordinary reactive attitudes. Thus, for example, we do not see resentment as *ever* constituting a fitting response towards the severely mentally handicapped. Despite these differences in scope, what happens in the case of both excuses and exemptions is that they render ordinary reactive attitudes inappropriate. Either because they show that the action is not indicative of the agent's attitude towards us (excuses) or because they reveal the agent to be generally incapable of expressing goodwill or regard towards us through his actions (exemptions). In both types of cases, the suspension of ordinary reactive attitudes is prompted

[17] Ibid., 9.
[18] Ibid., 7.
[19] Ibid., 8.
[20] Ibid., 7.

by the recognition of their being something special or abnormal about the action or the agent in question.

But *second,* as I said above, Strawson also offers an alternative way of approaching the reactive-objective distinction by providing a careful description of the behavioral patterns and cognitive dispositions characteristic of the objective attitude. These patterns and dispositions are summarized by Strawson as follows: "To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account of; to be managed or handled or cured or trained; perhaps simply to be avoided".[21]

Strawson contends that regarding other human beings in such ways is incompatible with displaying "the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships".[22] This second definition, however, conflicts with the first. This is because as Strawson himself says "we *can* sometimes look with something like the same eye on the behaviour of the normal and the mature[...] as a refuge, say, from the strains of involvement; or as an aid to policy".[23] In other words, valid excuses or exemptions are not necessary for it to be appropriate to occupy the objective attitude.

I believe we all have firsthand experience of taking the stance described by Strawson as the objective attitude. Everyone is familiar with coming to see another person "as a case".[24] Equally familiar is the decision to review one's assessment of an action upon learning that the agent could not have done otherwise. Such a shift in perspective frequently brings about a change in one's emotional reactions too, not unlike in the way described by Strawson. Gut reactions are checked, the immediate give-and-take of spontaneous human transactions is subjected to reflective scrutiny. Parents can often be seen to do this when dealing with their children. In many cases the mechanisms of emotional control and reflective monitoring of one's instantaneous reactions can even come to be institutionalized, most prominently perhaps between the psychologist and his patient in a therapy. But a similar process of reflection can also set in when it occurs to us that the other person's at first sight surprising or offensive behaviour is due only to cultural or social differences *and not* some serious handicap or deficiency. So although it is perhaps right that responsibility-undermining conditions give us sufficient reasons to suspend ordinary reactive attitudes, they do not seem to be necessary for us to do so.

But why is it a problem for Strawson's account that there may be good reasons to also occupy the objective attitude vis-à-vis, to use Strawson's

[21] Ibid., 9.

[22] Ibid.

[23] Ibid.

[24] Bennett, 'Accountability,' 24.

expressions, "the normal" and "the mature"? The difficulty is that once this concession is made there will no longer be a direct correlation between 'being an appropriate addressee of ascriptions of responsibility' and 'being an appropriate addressee of ordinary reactive attitudes'. The lack of such a correlation suggests that it is perfectly possible to occupy the objective attitude also when dealing with agents whose responsibility for their actions is undiminished. That is a significant finding because it undermines Strawson's attempt to draw the distinction between reactive attitudes and objective attitude in terms of when it is appropriate to hold someone responsible for his action and when it is not.

So we are still in search of a way of drawing the distinction between reactive attitudes and objective attitude which both provides a faithful account of situations in which we tend to suspend ordinary reactive attitudes and at the same time lends support to the Constitution Thesis according to which to hold one responsible just is to be susceptible to reactive attitudes.

Perhaps a more careful way of drawing the distinction will help. I will review three potential improvements on Strawson's original attempts. These proposals are reconstructed from suggestions to be found in the relevant literature on the reactive-objective dichotomy.

First, one may seek to characterize the objective attitude by arguing that objectivity is incompatible with feeling, empathy or personal involvement. Such a characterization is offered, for example, in the following statement: "A world in which human relationships are restricted to those that can be formed and supported in the absence of the reactive attitudes is a world of human isolation so cold and dreary that any but the most cynical must shudder at the idea of it."[25]

Wolf is of course right in highlighting the enormity of the loss we would be facing if we were to give up reactive attitudes altogether. But she is wrong to suggest that the absence of reactive attitudes is paramount to the absence of all feeling and emotion. None of the patterns of behavior or cognitive dispositions adduced by Strawson as typical of the objective attitude–i.e. regarding another person as an object of social policy or a subject of treatment or something to be managed, cured, trained, etc.– necessarily presupposes a lack of feeling or involvement. What's more, in many cases the success of relationships referred to as objective by Strawson positively require feelings and involvement.

Consider treatment. Therapists may well entertain various feelings towards their patients (and conversely). In fact, it is hard to see how the therapeutic process could be successful if interactions with the patient did not give rise to various emotions in the therapist himself. In addition, as Bennett rightly points out, therapist and patient are closely *involved* with

[25] Wolf, 'The Importance of Free Will,' 391.

one another and fully *participate* in the therapeutic effort.[26] Similarly, a politician may well have to take a step back from his immediate feelings and sympathies when designing social policies, but these policies are doomed to failure if they are not based on the right kind of involvement with the concerns of those whose interests they are supposed to serve. In any case, the lack of feeling is certainly not required for the success of those policies.

It is instructive to pause for a second here and to contrast the attitude(s) of the psychopath, on the one hand, with that of the therapist or politician or scientist or parent who goes for the objective attitude, on the other. It is the former who lacks feelings or at least lacks other-regarding feelings. The "cold and dreary world of human isolation" described by Wolf is that of the psychopath not of the therapist or the policy-maker (unless of course they themselves suffer from psychological disorders). It is psychopaths who "feel no *guilt, regret, shame* or *remorse"* and who "do not *care* about others or their duties to them [and] have no *concern* for others' rights and feelings".[27] It is, therefore, the psychopath who is "morally dead",[28] not those opting for the objective attitude for one reason or another.

I would suggest that what is special about the therapist is not the lack of feelings but what he does with them. In his role as a therapist he is expected to exercise control over his emotions and continuously subject them to careful scrutiny. This is because his task is to serve the interests of his patient. In doing so he is to set aside his own concerns and self-regarding feelings (such as whether he enjoys working with the patient for example).[29] Ideally, the same requirement should hold for makers of social policies.

However, not all the patterns of behavior or cognitive dispositions associated with the objective attitude by Strawson impose the same requirements as those shouldered by the therapist or the policy-maker. Thus for instance one may very well want to regard an overtly jealous friend "as a case", taking the objective attitude towards him when having to deal with this unpleasant trait, in order to protect one's own interests and welfare. Strawson himself stresses that we sometimes opt for the objective attitude towards other human beings in order to avoid them or to "take precautionary account of" them.[30] At this point, therefore, it seems that we can assume the objective attitude for many different reasons and can behave in quite divergent ways while remaining objective. Consequently, we still lack both a general characterization of the objective attitude and how it is supposed to be opposed to reactive attitudes.

[26] Bennett, 'Accountability,' 35.

[27] The description of the psychopath as morally dead is taken from Murphy's 'Moral Death: A Kantian Essay on Psychopathy,' 286-7.

[28] "Ohne alles moralische Gefühl ist kein Mensch; denn, bei völliger Unempfänglichkeit für diese Empfindung, wäre er sittlich tot." Kant, *Die Metaphysik der Sitten,* 531.

[29] See Stern, 'Freedom, Blame, and Moral Community,' 77.

[30] Strawson, 'Freedom and Resentment,' 9.

Second, there is A.J. Ayer's observation that "we tend to adopt an 'objective' rather than a 'personal' attitude towards a particular action, or towards the over-all behaviour of a particular type of agent [when] we do think that we command a set of scientific hypotheses from which, in conjunction with facts which are practically ascertainable if not already ascertained, the conclusion that the behaviour takes place can be derived with at least a high degree of probability and in quite a specific form, even if it does not reach down to every detail".[31]

It is certainly true that the therapist, for instance, relies on a set of theoretical hypotheses to understand and predict the behaviour of his patient as well as to guide and interpret his own reactions. The patient is regarded as a "case" (and as far as his participation in the therapeutic process is concerned the therapist must strive to regard himself as a "case" too). Similarly, when designing social policies decision-makers act on general assumptions and predictions about people's behaviour. Instead of turning to personal communication or direct means of persuasion they often rely on various measures calculated to exert an influence on people in line with the relevant policies.

But it is nevertheless questionable whether this observation can really be used to set apart the objective attitude from ordinary reactive attitudes. For one thing, drawing the distinction in these terms paints the wrong picture of ordinary reactive attitudes. In an effort to emphasize that reactive attitudes and feelings are naturally ingrained (a move which is crucial for the inescapability arguments to be discussed below), some passages in *Freedom and Resentment* may perhaps falsely lead us into thinking of such responses as being natural, spontaneous and immediate. But no matter how much they may be imbued with feeling, ordinary reactive attitudes can be and frequently are based on fine-grained perceptions and assessments of others. It is important to stress that a response being emotionally-charged does not exclude it being based on a complex set of beliefs and law-like generalizations (or even theory) as to how to interpret the given situation, what the other people involved in it are like, how they in turn will react to a certain response, what is the value and purpose of expressing the emotion, etc.

On the other hand, one may object, isn't there a perspective seen from which the behaviour of other people appears to resemble almost indistinguishably the workings of a complex mechanism? And isn't it true that certain reactive attitudes are incompatible with occupying this specific perspective? If from that perspective some behaviour appears to be deterministically governed by scientific laws, would we not from that perspective also question our right to feel resentful, angry or indignant about the given form of behaviour? And is it not this specific perspective Strawson refer us to when talking about the objective attitude?

[31] Ayer, 'Free-Will and Rationality,' 7.

Now, of course, Ayer may be quite right that a certain view of human behaviour as mechanistically determined is incompatible with some range of ordinary reactions to such behaviour. But this is not what Strawson has in mind when referring to the objective attitude. Strawson makes it quite clear, as I already quoted, that taking the objective attitude is prompted by the recognition of the presence of special circumstances or abnormalities (even when dealing with "the normal and the mature"). Taking the objective attitude is never based, he says, on viewing people's behaviour in general as being deterministically governed by scientific laws: "when the suspension of such an attitude or such attitudes occurs in a particular case, it is *never* the consequence of the belief that the piece of behaviour in question was determined in a sense that all behaviour *might be,* and, if determinism is true, all behaviour *is,* determined in that sense."[32]

In other words, the consideration 'he could not have done otherwise' that underlies responsibility-undermining excuses and exemptions is not of the same type as the consideration 'he could not have done otherwise *because* the general thesis of determinism is true'. The first type of consideration can move us to take the objective attitude. But this is not because it makes us view other people's behaviour as being deterministically governed by scientific laws.

In short, the problem with drawing the reactive-objective distinction on the basis of Ayer's suggestion is, first, that we would end up with a misleading picture of reactive attitudes as excluding well-founded generalizations and predictions about other people's behaviour, and second, that it would also contradict Strawson's own understanding of the reasons for assuming the objective stance which is not that when we are objective we regard human behaviour as describable in terms of scientific generalizations. Therefore, it is still open whether a general characterization of the reactive-objective distinction is attainable in a way that the characterization remains at the same time congruent with Strawson's brand of compatibilism.

The *third* proposal is quite modest seeking not to define reactiveness in general terms, but only to name one of its distinguishing, negative properties. However, I think it is the most helpful interpretation of the objective-reactive distinction. Jonathan Bennett suggests that what's common to all reactive attitudes is that they cannot "be the cause or explanation of *x*'s engaging in teleological inquiry".[33] According to Bennett's definition, a teleological inquiry aims at achieving some practical end. In the specific context we are concerned with, a teleological inquiry, in the pursuit of a certain practical end, aims at finding out how an other agent works or why he did what he did. The suggestion is that when we are engaged in such a teleological inquiry we occupy the objective attitude, but reactive attitudes themselves

[32] Strawson, 'Freedom and Resentment,' 18.

[33] Bennett, 'Accountability,' 38.

never prompt us to engage in teleological inquiry and never explain why we engage in teleological inquiry.

To understand what Bennett is getting at here, it may be helpful at this point to rehearse a criticism made in the previous chapter of the consequentialist account of reactive attitudes. There I said that one of the things that was wrong with consequentialism was that it gave a narrowly 'instrumentalist' explanation of reactive attitudes as 'instruments' serving to encourage or discourage certain types of action in the future. By the same token, Bennett's point seems to be here that ordinary reactive attitudes are incompatible with such an instrumentalist, calculating frame of mind. When we start contemplating "how someone works" in order to get him to do something or discourage him from doing something in the future, we distance ourselves from ordinary reactive attitudes and begin to move towards the objective attitude. Of course, even when occupying the objective attitude we sometimes give vent to our resentment if that appears to serve a desirable end. But the objective attitude is also compatible with pretense and feigned reactions.

The case of genuine reactive attitudes is different. We do not react to what someone has done to us or for us with resentment, gratitude, etc. *in order to* promote a practical end or serve a forward-looking purpose. Such reactive attitudes, as Strawson says, are responses to the attitude we think other people's actions express, i.e. good or ill-will towards us, quite independently from any forward-looking goal.

It is a different issue that reactive attitudes can of course also motivate. It is a different issue because the fact that a certain reaction or emotion quite reliably moves us to Φ does not mean that the occurrence of such reactions or emotions can be explained by, let alone justified by, our having to Φ or wanting to Φ or needing to Φ. Consider guilt. As I already said in the previous chapter, one does not feel guilt *in order to* make sure that one will not again do the thing one feels guilty about. But of course it is also true that if one does happen to feel guilt, then the feeling of guilt often functions as a powerful action-guiding motive not to do the same thing again.

Having accepted so much of Bennett's proposal, it is important to note that we have not found evidence yet that the objective attitude would necessarily exclude the entertaining of certain reactive emotions. What the objective attitude involves and reactiveness does not is the teleological frame of mind in the service of some practical end. Once in this frame of mind one could very well continue to entertain reactive sentiments of all kinds, I believe, even resentment, indignation or blame. What will be different is that in the objective attitude it will be the forward-looking practical end (e.g. therapy, policy-making, training, etc.) that will determine how one chooses to relate to one's sentiments. Again, what distinguishes the therapist's interactions with his patient is not what he feels but what he is supposed to do with those feelings.

But the example of the therapist-patient relationship can also help to dispel another potential misunderstanding concerning the objective attitude. Taking up Bennett's idea, I argued that the distinguishing mark of the objective attitude is the 'teleological frame of mind', i.e. the readiness to control one's reactions to others as the pursuit of a given practical end may happen to require. But such a focus on results is not incompatible with *reasoning* with the person or persons towards whom the objective attitude is adopted. Those adopting the objective attitude are often in a position of authority (e.g. parents, therapists, decision-makers, etc.) and therefore they may be in a position to tell their interlocutors what to do or what to think. But this is not the same as treating the other person as *mere* means incapable of responding to and acting upon reasons. As Lawrence Stern noted, there is a crucial difference between calculation and manipulation.[34] Both strategies "focus on results" and are as such associated with the objective attitude, but only manipulation involves treating the other person as *mere* means. If that is true, then only manipulation is incompatible with recognizing the other human being as a responsible agent and not the objective attitude as a whole.

Bennett's suggestion is nevertheless helpful in isolating a feature that reactive attitudes do not manifest and the objective attitude does, namely the absence of the 'teleological frame of mind' from ordinary reactive attitudes. But of course Bennett is also right in saying that we still have not found a positive characteristic holding together the wide spectrum of reactive attitudes as a class.

In any case, even if such a characteristic could be found, the above findings appear to jeopardize Strawson's account. Most importantly, it seems that reactive attitudes do not overlap with a view of others as responsible agents. It is true that the acceptance of an exemption, i.e. a view of the other as not being responsible agent, may constitute a sufficient reason to adopt the objective view towards him as someone who is to be managed, cured or avoided. It is also true that the acceptance of an excuse, i.e. a view of a specific action as one for which the agent bears no responsibility, may constitute a sufficient reason to adopt the objective attitude towards that specific action as something to be prevented or handled. But there may be countless other reasons to take the objective attitude towards other people or towards specific actions. In short, it is not true that taking the objective attitude towards a specific action or towards an agent overall necessarily entails the judgement that the agent is not responsible overall or not responsible for the specific action at hand. There is no psychological impossibility or logical contradiction involved in taking the objective attitude towards an agent and regarding him at the same time as fully responsible for his actions.

[34] Stern, 'Freedom, Blame, and Moral Community,' 74.

Why do these findings cause difficulties for Strawson's account? After all, we found no reason to question his claim that the recognition that some responsibility-undermining condition (excuse or exemption) applies in a given situation is indeed incompatible with certain reactive attitudes. For example, one's accepting that X 'could not have done otherwise but to Φ' and one's continuing resentment against X for having Φ-ed appear to conflict irresolvably. That is as we should expect. We should not give up Strawson's valuable insight that certain reactive attitudes are closely tied up with ascriptions of responsibility.

The question is only how best to conceptualize the nature of that tie. The above analysis of the objective-reactive distinction suggests that it is perfectly possible to take an 'objective distance' from our reactive attitudes without relinquishing our view of others as responsible agents. And that poses a serious challenge to the Constitution Thesis. If it is possible to suspend reactive attitudes while continuing to ascribe responsibility, it no longer seems to be the case that reactiveness is constitutive of moral responsibility or that to hold oneself responsible just is to be susceptible to such reactions. I hope to lend further support to that criticism in the following sections as well as discuss its implications for the other two main theses summarized at the end of the previous section.

4.4 Reactive attitudes and the "moral sentiments"

The difficulties raised in connection with the objective-reactive distinction cast doubt on Strawson's assertion that reactiveness is *constitutive* of responsibility (the Constitution Thesis). In this section, I will try to question the somewhat weaker, but still very strong claim that reactive attitudes are the key to understanding responsibility. I think, in the end, we should rest content with the even less ambitious proposal that reactive attitudes are only constituent parts of responsibility-attributing practices. Very important parts, but only parts nevertheless, which in many cases may even fall to the wayside. To have shown this much already qualifies Strawson's conception as a supremely important contribution to the theory of responsibility.

According to the suggested redrawing of the objective-reactive distinction in the previous section it is possible to view another person as a responsible agent while not manifesting ordinary reactive attitudes such as resentment or indignation. One significant implication of that conclusion is that an ascription of responsibility is not *necessarily* an affective phenomenon. What needs to be shown now is that an ascription of responsibility is not even *primarily* an affective phenomenon. In terms of one of the distinctions (emotional response vs cognitive content) constituting the 2x2 classificatory matrix introduced at the beginning of this chapter that would mean that the cognitive content, i.e. a judgement of responsibility, is prior to the reactive emotions which, undoubtedly, ascriptions of responsibility may (but need not) involve. And in terms of the other distinction constituting that matrix (holding vs being responsible) it would mean that being responsible is prior to holding responsible.[35]

I will try to show what those priority-claims entail by focusing in this section on yet another distinction put forward by Strawson, namely the distinction between what Strawson calls 'moral reactive attitudes' from their 'non-moral' counterparts. As will be seen below, it is not entirely clear what is at stake for Strawsonians in making this distinction. Why should it matter that certain reactive sentiments are moral? In what ways does the use of this adjective change the significance of these reactive sentiments? These questions are not really addressed in *Freedom and Resentment.* At the same time, the distinction is relevant for the purposes of this work because those reactive sentiments which Strawson classifies as moral (e.g. resentment, guilt) are precisely those which have been identified in Chapter 2 as requiring an ascription of responsibility.[36] So whatever Strawson's intention may have been in making this distinction, it remains an important question whether it is possible and on what grounds it is possible to separate these two types of reactive sentiments. My proposed conclusion is that this question is not satisfyingly answered by the Strawsonian theory.

As a number of commentators noted, and Strawson himself later admitted, the way this distinction came initially to be drawn in *Freedom and Resentment* is in any case misguided. But, as I said, I want to venture a bolder claim here. What I would like to argue in this section is that there is simply no tenable way of drawing that distinction without conceding that judgements of responsibility and being responsible are prior to emotional responses and holding responsible, respectively. In sum, there is no way of drawing a line between these two types of reactive attitudes other than in terms of some form of judgement concerning the agent's responsibility.

But, first, let us retrace our steps a little. We have found no reason in the foregoing to question Strawson's claim that all reactive attitudes necessarily involve seeing the other agent (and if the reactive attitude is reflexive such as guilt for example is, then seeing oneself) as capable of expressing or being in a position to express goodwill. Reactive attitudes are a response to the 'quality of will' manifest in the actions of others. If the original action expresses a lack of goodwill, then the response too may involve a withdrawal of goodwill as expressed in characteristic forms of indignant, resentful, etc. behaviour. On the whole, therefore, my taking a reactive stance towards someone means that I recognize his agency. Depending on how that agency manifests itself towards me I may respond to it one way or another.

[35] The combination of these claims was called the Priority Thesis in Chapter 3, see p. 42 and below.

[36] See Chapter 2, p. 35f.

These reactive attitudes are part and parcel of interpersonal relationships as Strawson makes it clear. Note, however, that not all interpersonal attitudes and emotions are necessarily reactive. Consider disgust. One can feel disgusted by some object as well as a person. The feeling of disgust is not tied up with a view of the object of disgust as an agent on its own right. The same goes for lust (whereby I mean a combination of arousal and affection). In fact, perhaps some would argue that not even all forms of love require regarding the object of love as an agent.[37]

In any case, a far more pressing question is why certain reactive attitudes appear to presuppose an ascription of responsibility while others do not? This is the central and most problematic distinction to be discussed in this section. As I said, it is never really explained in *Freedom and Resentment* what the implications of making that distinction would be.[38] That is, suppose we did find a satisfying way of separating so called 'moral' reactive attitudes from 'non-moral' ones, what would then follow from the fact that a certain reactive attitude is 'moral'? Although, as already mentioned, Strawson never really answers that question, it is reasonable to assume that the distinction is necessary because we have different reasons for the type of response that Strawson classifies as moral. For example, we have different reasons for feeling anger (assumed to be a non-moral reactive attitude) than for feeling resentment (assumed to be a moral reactive attitude). My claim then is that not only is there no satisfactory way of accounting for that difference, but that it is in principle impossible to draw that distinction within the framework of the Strawsonian theory.

In any case, Strawson himself draws the distinction as follows. Moral reactive attitudes are those which are "reactions to the qualities of others' wills, not towards ourselves, but towards others. Because of this impersonal or vicarious character, we give them different names. Thus one who experiences the vicarious analogue of resentment is said to be indignant or disapproving, or morally indignant or disapproving."[39]

But this definition cannot be right. Vicariousness cannot be what distinguishes moral reactive attitudes from others. First, as Bennett notes (a criticism which Strawson later accepted), this definition would unacceptably

[37] As Jay Wallace rightly notes, only by equating reactive attitudes with involvement in interpersonal relationships can Strawson lend credence to his claim that reactive attitudes are natural and inescapable for normal human beings, see Wallace, *Responsibility and the Moral Sentiments,* 31. But this equivocation, as we now see, is not persuasive which in turn casts doubt on the Inescapability Thesis as well. Strawson of course presents other arguments too in favour of the Inescapability Thesis. These will be discussed in the final section of this chapter.

[38] This recalls Williams's vigorous criticism of a pattern of thought according to which the distinction between the moral and the nonmoral is "at once deep, important and self-explanatory", Williams, *Shame and Necessity,* 92.

[39] Strawson, 'Freedom and Resentment,' 14.

classify reflexive attitudes, such as guilt or compunction, as non-moral.[40] Second, as Jay Wallace notes (but Bennett does not), this definition would also exclude feeling *moral* resentment on one's own behalf, say, about being treated unfairly.[41]

It seems, however, that these are omissions not easily rectifiable within the framework of the original Strawsonian account. Consider the case of reflexive attitudes first. Clearly, my feeling of guilt is a reaction to what I have done (to someone else). Yet it is hard to see how the normative relevance of guilt, the action-guiding force of feeling this emotion, can be accounted for in terms of responses to the agent's 'quality of will'. Why should I care about the 'quality of my own will' towards others? Perhaps because I fear the withdrawal of their goodwill. But this need not happen and yet I can feel genuine guilt. Indeed it seems that the real reason why I care about the quality of my own will towards another human being is because I believe (rightly or wrongly) to have violated a standard, norm or expectation.

And if that is true, then we may begin to wonder, second, why we care about the quality of others' will towards ourselves. Do we only react with resentment in such cases because our "own interest and dignity"[42] is at stake? Or is the feeling of resentment caused not only by the setback to our interests or the injury suffered, but at least in part also by the belief that a standard, norm or expectation was violated in how we were treated? That this may be a more adequate understanding after all is shown by all those cases in which even though the other person manifests goodwill towards us, we nevertheless feel resentment because the action involves such a violation of some norm, standard or expectation.

The failure of Strawson's own attempt to account for the special significance of certain reactive attitudes in terms of their vicarious quality leaves us looking for an alternative explanation. In particular, we are still in search of the best way of conceptualizing the tie between certain reactive attitudes and ascriptions of responsibility. It is questionable, however, whether these tasks can be successfully tackled within the framework of the Strawsonian 'quality of will' theory.

The main reason for my saying so is that one important strength of that theory is also its major weakness. Thus the Strawsonian account paints a vivid picture of how closely and deeply emotions and attitudes and patterns of behaviour *and* beliefs and views of other human beings as agents are interconnected. It tells us that what holds that "complicated web of atti-

[40] Bennett, 'Accountability,' 46. Strawson accepts this criticism in his 'Reply to Ayer and Bennett,' 266.

[41] Wallace, *Responsibility and the Moral Sentiments,* 35.

[42] Strawson, 'Freedom and Resentment,' 14.

tudes and feelings"[43] together is a "basic demand for reasonable regard"[44] and goodwill to be shown by human beings in their transactions with one another. But the fact that there is a continuity at one level (i.e. all reactive attitudes involve the recognition of other people's agency) does not entail that there is no conceptual distinction to be made at another level. The 'quality of will' account exposes the continuum of reactive attitudes ranging from responses to personal injury all the way to moral indignation provoked by events far away or in the distant past. And at one level positing such a continuum is right and is revealing for all such reactions involve a recognition of the agency of others as well as the demand to be recognized as such an agent by others.

But at another level–as already foreshadowed by the above criticisms of Bennett and Jay Wallace–some forms of reactiveness involve something more and also something different from this basic demand. In other words, while it is true that all forms of reactiveness are motivated by an interest in other agents' attitude towards us, it is not true that the interest that motivates ascriptions of responsibility to other agents for their attitudes is reducible to a demand for goodwill by those agents.

The most important evidence that reactive attitudes which involve an ascription of responsibility are to be treated as a separate class is in fact offered already in *Freedom and Resentment*. Strawson's preoccupation with responsibility-undermining excuses or exemptions clearly reveals his awareness that it is not only the 'quality of will' displayed by the other person that we care about when ascribing responsibility. As Gary Watson puts it: "a child can be malicious, a psychotic can be hostile, a sociopath indifferent, person under great strain can be rude, a woman or man 'unfortunate in formative circumstances' can be cruel"[45] and yet, as Strawson is well aware, despite the negative quality of will that comes to the fore in such instances, we nevertheless tend to suspend our ordinary reactive attitudes.

It is true that the tendency to do so is to some extent just another basic empirical fact about the human psychology of reactiveness. This is most obvious in the case of children. Normal adults appear hard-wired not to react to children in certain ways. That is to say, we seem to be *naturally* prone to withhold some of our ordinary reactive attitudes–e.g. resentment, moral outrage–in our dealings with children.

But no such natural proclivity is observable in the case of criminal offenders even if they can be shown to have been the victims of early deprivation or to suffer from a severe mental handicap. Quite the contrary. In such cases, most people literally have to turn to 'their better selves' to 'fight off' their moral outrage and retributive instincts especially if the crime committed is

[43] Ibid., 23.

[44] Watson, 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,' 259.

[45] Watson, 'Responsibility and the Limits of Evil,' 262-3.

sufficiently serious. So clearly, the recognition of excuses and exemptions as incompatible with certain reactive attitudes is based in many cases on normative considerations rather than allegedly hard-wired psychological inclinations.

The discussion of the objective-reactive distinction in the previous section has already shown that reactiveness and ascriptions of responsibility do not necessarily go hand in hand. More specifically, already that discussion has shown that we suspend reactive attitudes not due to a psychological impossibility of resenting what, say, a psychotic person has done to us. Rather, we do so because we have good reasons to think that a given response is *inappropriate.*[46] Moreover, in the case of excuses and exemptions, we suspend ordinary reactive attitudes because we judge it to be unfair or cruel or unreasonable to impose certain normative consequences on people who couldn't help doing what they did.

In sum, what the relevance of excuses and exemptions to the attributability of responsibility shows is that ascriptions of responsibility involve not just reactions to the manifestations of goodwill or the lack of it on the part of other agents, but also substantial claims–or indeed judgements–about the addressee of the ascription. Unlike in the case of other reactive attitudes, those judgements are prior to potential ways of expressing them whether through emotional reactions or otherwise.

To avoid misunderstandings, I am not claiming that Strawson's theory is an emotivist one. That that is not the case is already shown by Strawson's meticulous discussion of the various responsibility-undermining conditions which can lead us to suspend our ordinary reactions. And he would of course quite happily second to Jeffrie Murphy's assertion too that "guilt, for example, cannot be identified as a feeling and distinguished from other feelings solely in terms of how, subjectively, it *feels."*[47]

In general, Strawson insists that reactive attitudes are not merely subjective emotional responses, but are *merited* by manifestations of good or ill-will. Since they are dependent on "a background of beliefs about the objects of those attitudes",[48] they can be subject to correction, modification and redirection.[49] In fact, Strawson positively requires that reactive attitudes be justified.[50] What I do wish to claim, however, is that the 'quality of will' theory does not have the resources to explain why and how those background beliefs are relevant to the justification of reactive attitudes. As

[46] Scanlon, 'The Significance of Choice,' 162-3.

[47] Murphy, 'Moral Death,' 287n8.

[48] Watson, 'Responsibility and the Limits of Evil,' 263.

[49] See for example Strawson, 'Freedom and Resentment,' 23.

[50] By proposing to read Strawson in this way I diverge from Bennett's gloss of the Strawsonian conception of reactive attitudes as expressing merely "my emotional make-up, rather than reflecting my ability to recognize a blame-meriting person when I see one". See his 'Accountability,' 24.

Gary Watson succinctly puts it: ”The problem is not just that the theory is incomplete, but that what might be necessary to complete it will undermine the theory”.[51]

What the Strawsonian theory seems unable to account for is why certain responses presuppose an ascription of responsibility, while others do not. As we have seen, all reactive attitudes are predicated on the notion of merit in the sense that all reactive attitudes involve ordered response-object pairs. For example, somebody's tactless remark may rightly anger me. But it would be inappropriate for me to feel insulted by that person if I think that the other person is not responsible for making that tactless remark because he is (say) unaware of the remark being tactless in the given context. Resentment would only be appropriate on my part if I judge the other person to be responsible as well. It is that difference that the 'quality of will' theory cannot account for.

I contend that the only way to account for that difference is by acknowledging the priority of cognitive content in the explanation and justification of reactive attitudes. This holds true for both other-regarding and reflexive responses. We have just seen that what gives us reason to feel resentment towards an agent is the ascription of responsibility to that agent. Similarly for reflexive responses: the agent does not have a reason to feel guilty for his action unless he is responsible for his action.

Putting the matter thus can also help to clarify the relationship between 'being responsible' and 'holding responsible' which remains at best ambiguous in the original Strawsonian account. Thus it is a frequently voiced complaint against Strawson's expressivist theory that it fails to take into account that any reaction to other people's actions can be kept ”strictly private”.[52] However, I do not think that anything Strawson says commits him to denying this. As we have seen, what he does say is that if the original action expresses a lack of goodwill, then the response too *may* involve a withdrawal of goodwill as expressed in characteristic forms of indignant, resentful, etc. behaviour. So Strawson's account, as Bennett notes, has ”no imperatives demanding indignation or any other reactive feeling, but only imperatives forbidding them in certain areas, and permissions to have them in the remaining areas.”[53] But his saying that reactive attitudes are *permissible* under certain conditions (i.e. when responsibility-undermining conditions, excuses or exemptions, do not obtain) is perfectly compatible with the possibility of keeping them private.

[51] Watson, 'Responsibility and the Limits of Evil,' 263. For the same reason, George Sher says the following: ”Strawson's own argument commits him to denying that anyone is blameworthy at a deep level.” Sher, *In Praise of Blame,* 81.

[52] Sher, *In Praise of Blame,* 86. For the same point see also Scanlon, 'The Significance of Choice,' 165-6.

[53] Bennett, 'Accountability,' 24.

What's more, Strawsonians (as well as consequentialists as we saw in the previous chapter)[54] could quite rightly press the question: what is the *point* of ascriptions of responsibility if they are kept private? It is alright to claim that in some circumstances it may not be opportune or tactful or productive to express an ascription of responsibility in the form of an overt reaction, whether emotionally-charged or otherwise. But aren't such withheld ascriptions parasitic on or at best side effects of the continuous overt exchange of approvals and disapprovals, criticisms and defenses, addresses and rebuttals, emotional give-and-takes? Don't responsibility-attributions constitute first and foremost a public practice whereby the internalized norms and patterns of that public practice will govern privately made ascriptions too which, therefore, can play at best only a derivative, secondary role? In short, doesn't the judgement underlying ascriptions of responsibility principally derive its special normative force from the fact that it *could* be public, or more precisely, that ultimately it refers us back to a public practice of the community?

But Strawson only tells us in a permissive vein that if an action expresses a lack of goodwill, then the response too *may* involve a withdrawal of goodwill. That 'may' calls for clarification and I doubt whether the Strawsonian theory has the resources to do so. The contrary proposal defended in this work is that 'being responsible' is prior to 'holding responsible' in the sense that one necessary condition for the overt expression of a reactive attitude to be *permissible* is that agent to whom that action is ascribed is indeed responsible for it. The agent incurs some normative consequences because of his action if and only if he is responsible for it. At the same time, 'being responsible' is a necessary but not a sufficient condition for the justification of overt responses. As we have already seen in other chapters, there may be a number of reasons (e.g. lack of authority, the person has already suffered enough, the response may be counterproductive, etc.) why, although the judgement of responsibility seems justified, the imposition of certain normative consequences (e.g. punishment, overtly expressed blame or resentment) is not.

But, to repeat, it is questionable whether such an answer can be accommodated within the framework of the original Strawsonian account. By regarding reactive attitudes which are kept private as at best parasitic on their publicly expressed counterparts, the Strawsonian theory lacks the resources to account for the fact that norms for ascribing responsibility are different from the norms for the overt expressibility of such ascriptions. This a point already made in the previous two chapters: justifying an ascription of responsibility is one thing, justifying the actual imposition of normative consequences (ranging from emotional reactions all the way to sanctions such as punishment) on an agent who is found to be responsible is another.

[54] See p. 51.

It is one of the principal shortcomings of the Strawsonian theory that it remains silent on this basic difference.

Unsurprisingly, the debate hardly comes to a close with this. After a brief digression on a special argument made by Strawson, in the final section of this chapter, I will discuss possible (but ultimately, I believe, unsuccessful) retorts to the cognitivist arguments made here.

4.5 Goodwill vs. freewill: failure of the generalization strategy?

The discussion to be undertaken in this section may seem at first sight to constitute something of a digression. It takes up the issue, or at least Strawson's approach to it, whether determinism is compatible with moral responsibility. My working method so far has been to assume, without further inquiry, that either some form of compatibilism or libertarianism is correct. This method was adopted in the hope that, first, it makes it possible to avoid at least to some extent the quagmire of questions concerning the metaphysical pre-conditions of responsibility, and second, that there is much to be said about the normativity of responsibility-ascriptions beyond those metaphysical questions too. In this section, however, I will diverge somewhat from this working method. This is because Strawson's arguments in favour of his own special brand of compatibilism are directly tied up with his explanation of *why* certain conditions–the responsibility-undermining excuses and exemptions already discussed at length–make it inappropriate to ascribe responsibility. In short, the Strawsonian defense of compatibilism is closely linked to a specific understanding of how ascriptions of responsibility are to be justified. For this reason, the findings of this section will also bear closely on the problems to be raised in the following one as well as in Chapters 5 and 6.

Specifically, this Strawsonian brand of compatibilism is based on a denial of the "generalization strategy".[55] That strategy is a favorite among incompatibilists because it promises to demonstrate not only that the truth of determinism and responsibility cannot be reconciled, but also that the acceptance of incompatibility is forced upon us by our most basic intuitions concerning the attributability-conditions of responsibility. Thus it is argued that the considerations which lead us to absolve agents from responsibility for their actions (locally in the case of excuses and globally in the case of exemptions) are such that they would generalize for all actions and all agents if determinism was true. If the thesis of determinism is true all our actions are, so the incompatibilist intuition, like those of children or epileptics in the relevant respect (and it is up to the incompatibilist to fix this

[55] The term itself is adopted from Wallace, *Responsibility and the Moral Sentiments,* 16-7 and passim.

relevant respect in terms of lack of control or lack of alternate possibilities or whatever he takes to be that necessary condition of responsibility which is vitiated by determinism). Determinism, in short, would operate as a blanket responsibility-undermining condition and would make it the case that no one would ever be responsible for anything. Or conversely, the truth of determinism would rule out that the conditions required for the attribution of moral responsibility could ever be met.

Most typically, the incompatibilist's claim is that we believe that for the agent to be responsible for what he did it must be true that he could have done otherwise (the principle of alternate possibilities). We admit excuses (e.g. coercion, etc.) and exemptions (e.g. serious mental impairment, etc.) as relevant considerations precisely because they entail that the agent could have not done otherwise. If, however, determinism also entails that the agent could not have done otherwise, then determinism entails that no one is ever responsible for anything.

Those who deny the generalization strategy object that the reasons why we excuse or exemption agents from responsibility have nothing to do with the thesis of determinism. In Strawson's idiom this translates into the claim that when we do suspend reactive attitudes in a particular case, this is "*never* the consequence of the belief that the piece of behaviour in question was determined in a sense that all such behaviour *might be,* and if determinism is true, all behaviour *is,* determined in that sense."[56] As we have seen, on Strawson's account, excuses and exemptions block the withdrawal of goodwill in response to the agent's action because they show that that action should not be construed as expressive of the agent's original attitude towards us. This is either because the agent was temporarily not in a position to express his attitude (e.g. he acted accidentally, under compulsion, etc.) or lacks the general capacity (e.g. due to being morally underdeveloped or psychologically abnormal) and is therefore an inappropriate target of certain reactions. But, so Strawson, it cannot be true that these special conditions hold for all agents and all actions.

One great advantage of this way of arguing in favour of compatibilism seems to be that it frees one from the burden of having to spell out the exact meaning of the thesis of determinism itself. Whether determinism entails the lack of alternate possibilities or something else, the theoretical commitment to it is unconnected to our acceptance of excuses and exemptions as relevant responsibility-undermining conditions. So the Irrelevance Thesis[57] can be read as saying that whatever one's exact definition of determinism may be,

[56] Strawson, 'Freedom and Resentment,' 18.
[57] See Section 4.2, p. 80.

as long as this definition is to be taken as extending to *all* agents and *all* action, it cannot be relevant to the attributability of moral responsibility.[58]

In *Freedom and Resentment,* Strawson puts forward two arguments in support of a denial of the generalization strategy along these lines. These are contained in the summary overview of Strawson's reconciliatory compatibilism in Section 4.2 under Step 3 and Step 4. I would like to discuss these two arguments now in more detail.

I begin with Step 3. Strawson contends that the objective attitude can only be assumed "temporarily" and never for very long.[59] Universal and thoroughgoing detachment is impossible because human beings are psychologically incapable of assuming the objective stance as a default position. In addition, doing so would involve abandoning the web of interpersonal relationships and the price to be paid for this would be to lose one's humanity. No doubt, we can, and indeed we should, assume the objective attitude in those special cases when excuses and exemptions apply. However, accepting determinism would require us to occupy the objective attitude as the default perspective towards all agents and all actions. And that option is simply not available to human beings.

One problem with the *argument from the impossibility of a universally objective attitude* is that once the objective-reactive distinction is subjected to closer scrutiny the argument begins to lose some of its original appeal. Once it turns out, first, that objectivity is not inconsistent with seeing other human beings as responsible agents (as I argued in Section 4.3), and second, that "reactive attitudes are not coextensive with the emotions one feels toward people with whom one has interpersonal relationships",[60] assuming a universally objective stance, although still possibly not a very attractive choice, certainly no longer appears to be impossible or wholly inimical to human nature.

Setting these difficulties aside for the moment it should also be repeated that there is also a tension between this argument and Strawson's own fine-tuned analysis of reactive attitudes.[61] Thus he readily concedes that we can for various reasons suspend reactive attitudes also when facing someone entirely mature and normal. In fact, he repeats the point originally made in *Freedom and Resentment*[62] with added emphasis in a later work: "I mean that there is open to us the possibility of having deliberate recourse to an objective attitude in perfectly normal cases; that it is a resource we can

[58] Ibid., 1, 10, etc. As I already indicated (see p. 80n15), the Irrelevance Thesis can also be reached indirectly via the Inescapability Thesis. This option will be discussed in the following section.

[59] Ibid., 10, 12.

[60] Wallace, *Responsibility and the Moral Sentiments,* 31.

[61] See Section 4.2, p. 77f.

[62] Strawson, 'Freedom and Resentment,' 9-10.

sometimes temporarily make use of, for reasons of policy or curiosity or emotional self-defense."[63]

But if that is true, one can rightly wonder what exactly is the reason why we could not occupy the objective attitude universally towards all agents and all actions. More precisely, we need some additional argument to explain why we cannot "hold, or rest in, for very long"[64] in the objective attitude towards "the normal" and "the mature". Without such an argument it will be difficult to explain why it could not be a realistic consequence of accepting the truth of determinism that the objective attitude becomes, if perhaps not our exclusive, but at least our default perspective on human behaviour.

Can the second argument, the *argument from self-contradiction* fill this explanatory gap? With this question I now come to Step 4 in Section 4.2. Strawson's idea here is that the attempt to generalize the considerations which lead us to accept excuses and exemptions in particular cases, and as a result suspend our reactive attitudes, is self-contradictory. This argument is based on Strawson's claim, which I have already discussed in connection with the reactive-objective distinction in Section 4.3, that every time we suspend our reactive attitudes this is because we recognize that there is something abnormal either about the agent or something exceptional about the situation in which the agent acted. For instance, the injury may have been caused accidentally ("he didn't mean to", "he was pushed") or the agent was non-culpably ignorant of potential consequences of his action ("he hadn't realized") or acted under compulsion or duress ("he had to do it").[65] Alternatively, we realize that the agent is special in the sense that he is not quite an agent in the full sense of the word on account of being "morally underdeveloped" or "psychologically abnormal" or because "he wasn't himself" having been brainwashed, or hypnotized or exposed to great stress.[66]

Consequently, such excuses and exemptions apply to particular cases precisely by virtue of the fact that these cases diverge from the normal and the ordinary. The thesis of determinism states, by contrast, that such special conditions apply universally to all agents and all actions. But that thesis then is logically self-contradictory because it implies that "abnormality is a universal condition".[67] No matter what the precise meaning of the thesis of determinism may be and quite apart from the psychological reality of accepting this thesis, we know that it seeks to extend considerations–which can obtain by definition only in a limited number of instances–to all agents and all actions. That alone is sufficient to render the thesis self-contradictory.

[63] Strawson, *Skepticism and Naturalism. Some Varieties,* 34.

[64] Ibid.

[65] Ibid., 7.

[66] Ibid., 8.

[67] Ibid., 11.

I believe, however, that Paul Russell is quite right to point out that this argument is misguided as it depends on an illicit "conflation or equivocation between being 'abnormal' and 'incapacitated'". And, as Russell continues, "[if] we replace Strawson's references to 'the abnormal' and 'abnormality' with references to 'the incapacitated' and 'incapacity' [Strawson's] reply to the Pessimist [i.e. the incompatibilist], quite simply, collapses".[68] Strawson's argument collapses because there is of course no self-contradiction involved in assuming that all agents are incapacitated in the relevant respect, namely in the respect of lacking responsibility-entailing freedom. In fact, this is precisely what many incompatibilists take the thesis of determinism to entail.

The upshot is that there is no 'quick fix' for those who seek to defend compatibilism by undermining the generalization strategy. That said, there are other more forceful arguments (not mentioned by Strawson) available to compatibilists of this persuasion. One further possibility to block the generalization from excuses/exemptions to determinism as a blanket responsibility-undermining condition is to deny that a unified account account of the considerations or principles underlying excuses/exemptions can be given. This would still leave it open for incompatibilists to argue that (unconnected to excuses/exemptions) the truth of determinism is to be recognized as an independent responsibility-undermining condition. But this argument could no longer exploit the intuition that determinism is pertinent to the attributability of responsibility *for the same reason* that excuses/exemptions are relevant to the attributability of responsibility, thereby depriving the incompatibilist strategy of some its original intuitive appeal.

Another even more promising candidate to block the generalization strategy is spelled out with great care and precision by Jay Wallace. He argues that the principle underlying standard excuses/exemptions is not the principle of alternate possibilities, as is most often assumed, but rather quite different principles of fairness. Wallace maintains that these principles, while better to suited to explaining why we recognize standard excuses and exemptions than the principle of alternate possibilities, would not generalize even if determinism turned out to be true.[69] He argues that excuses function not because they indicate that the agent could not have done otherwise, but rather "by showing that an agent has not really done anything wrong". From that it would follow that the moral force of excuses is to be accounted for by the principle of "no blameworthiness without fault" rather than the principle of alternate possibilities.[70] Exemptions in turn are based on the consideration that in the relevant cases agents lacked "powers of reflective self-control" (roughly, the ability to grasp and act in accordance with moral

[68] Russell, 'Strawson's Way of Naturalizing Responsibility,' 299.

[69] Wallace, *Responsibility and the Moral Sentiments,* 115-6.

[70] Ibid., 135.

reasons). Once again, this ability does not presuppose the ability to do otherwise, and what's more, it is not incompatible with the truth of determinism. Therefore, determinism would not entail a generalization of the standard exemptions for all agents and all actions.[71] Or so Wallace argues.

I will not try to assess here the merits of Wallace's proposal since I believe that even if it does succeed in blocking the generalization strategy it will not be able to fully allay the incompatibilist's worries. This is because those who base the defense of their compatibilist position on the denial of the generalization strategy falsely assume that determinism yields no independent normative considerations.

It is indeed a great merit of the Strawsonian approach to have questioned our faith in the viability of the generalization strategy. The incompatibilist's attempt to identify a basic feature common to all responsibility-undermining conditions, whether local or global, is not without its own difficulties. It seems even more problematic to establish that if determinism were true this feature would obtain for all agents and all actions.

I would disagree, however, with the claim that determinism is simply irrelevant as a normative consideration. Irrespective of the fate of the generalization strategy, what is at issue is the justifiability of our responsibility-attributing attitudes or judgements. This issue is, however, essential because we naturally seek to make judgements concerning the responsibility of this or that agent *which are true.* Moreover, we expect these judgements not only to be true but to be be true in a non-contingent fashion. That is precisely the reason why we are interested in justification for what justification does is to increase the likelihood of a belief being true.[72] In other words, our responsibility-attributing judgements and attitudes should be backed up by reasons.

The interest in the truth of our responsibility-attributing judgements could in itself be significant enough to recognize the normative relevance of determinism. But our readiness to acknowledge the normative relevance of determinism will only increase once we realize that this interest may not be merely driven by the self-contained pursuit of truth, but also by the consideration that as long as we cannot guarantee the truth of responsibility-attributing judgements we cannot guarantee that they remain appropriate or fair either. The essential worry is that if the truth of determinism gives us reason to doubt that our judgements of responsibility will be true, then it also gives us reasons to doubt that they will be appropriate and fair. If that is correct, then the normativity of determinism will be not only theoretical but also practical.[73]

[71] Ibid., 155.

[72] See Farkas, 'Szkepticizmus és Filozófiai Gondolkodás,' 61.

[73] At the end of the following section, I will criticize the Kantian approach for not giving sufficient weight to this consideration.

Now nothing said so far gives us reason to doubt that a compatibilist analysis of the traditional kind could succeed in showing that the truth (or falsity) of determinism does not necessarily render attributions of responsibility unjustifiable. But this cannot be done by circumventing the problem of determinism as reconciliatory compatibilism would want us to do. I will now turn to Strawson's principal argument–the inescapability of responsibility (which, as we will see, is a bundle of arguments rather than a single one)–in favour of reconciliatory compatibilism.

4.6 Does inescapability justify? Three arguments

The argument from inescapability is in many ways the centerpiece of Strawson's theory of moral responsibility. The theme of inescapability resurfaces in other areas of Strawson's work too, most prominently in his proposed neo-Humean solution to skepticism.[74] Strawson's interest in inescapable beliefs and practices and his preoccupation with how such inescapability may impact on the justification of these beliefs and practices is often referred to as Strawsonian naturalism.[75] Strawson himself is partly responsible for coming to be labelled in this way. After all, he himself explicitly refers to his preferred way of meeting different forms of the skeptical challenge in epistemology, morality and elsewhere as "non-reductive naturalism".[76] However, I would like to argue that the argument from naturalism is just one way for Strawson to articulate and highlight the importance of a more comprehensive concern with 'inescapability' and, specifically, the connection between 'inescapability' and justification. I will try to show in the following that Strawson advances a number of separate and to some extent conflicting arguments to support his principal claim that–taking some liberties with the depth and sophistication of Strawson's insights–one could reduce to the dictum: 'inescapability justifies'.

Bringing to mind the explicit analogies drawn between Strawson's proposed treatment of skepticism and his reconciliatory compatibilism can help us to see that justification lies at the heart of the whole inescapability issue. The essence of Strawson's solution to the skeptical challenge in epistemology is that skeptical arguments are to be *circumvented* rather than directly answered.[77] We can do so once we realize that our commitment to certain

[74] See Strawson, *Skepticism and Naturalism. Some Varieties.* The words 'inescapable' and 'inescapability' or synonymous expressions occur with remarkable frequency in this work. The connection between Hume's thought and Strawson is explored among others in Williams, *Unnatural Doubts,* esp. xiii., 11-5, 24-5, etc.

[75] For example, Sher talks of Strawson's "uncompromising naturalism" and describes his theory as "relentlessly naturalistic". See Sher, *In Praise of Blame,* 81, 85.

[76] Strawson, *Skepticism and Naturalism,* 24, 39-41.

[77] See esp. ibid., 3. Cognate forms of the indirect solution to skepticism have been put forward by Hume, Wittgenstein, Carnap, Heidegger, or more recently, by Barry Stroud and Michael Williams. For an insightful and critical discussion of indirect responses to

beliefs is inescapable. These are beliefs that we "cannot help accepting", which we "take for granted", which we "neither choose nor could give up".[78] What that means is that we neither have nor *need* to have reasons for which we hold these beliefs. But if that is true, then the skeptic's demand for a justification of such beliefs is "idle", "unreal", "inefficacious", altogether besides the point, or even, on a stronger version of the argument, incoherent and meaningless.

One important feature of such beliefs which we are alleged to be inescapably committed to is that they define the framework of our thinking about an entire area. Thus the belief in the existence of the physical world is definitive of our thinking about particular objects, the belief in the general reliability of induction is definitive of how we arrive at specific empirical generalizations, the belief in the reality and determinateness of the past is definitive of our thinking about particular historical facts. These framework-beliefs provide the "scaffolding", "background", "substratum" (this is Strawson quoting Wittgenstein's pertaining metaphors with approval)[79] of entire belief-systems and practices.

Now, it is not entirely clear at this point whether the indirect anti-skeptical argument rehearsed here is that (i) these framework-beliefs require no justification *full stop,* or (ii) they require no justification because they are maximally general, as it were, so that we cannot come up with even more general reasons why we should hold them, or (iii) they are justified by virtue of their 'enabling-role', i.e. our commitment to them is justified by the fact that they make it possible for us to conceptualize and form specific beliefs within the framework which they come to define.[80] No doubt, option (iii), by producing something like a 'reverse justification' for framework-beliefs, is in some tension with the claim that there is no reason why we hold framework-beliefs at the first place. In any case, I believe that the ambiguity indicated here never gets to be resolved in Strawson's proposal and is also reflected in the discrepancies surfacing between his various arguments from inescapability in the area of responsibility-ascriptions to which I now turn.

skepticism, see Farkas, 'Szkepticizmus és Filozófiai Gondolkodás,' esp. 68-70 and 72-4. I find myself in disagreement, however, with the contention of that article (cf. 71) that all those who propose an indirect response to skepticism base their solution on a distinction between everyday language and ideas, on the one hand, and philosophical thought, on the other. This may be true of Hume, for example, but I think it is not the dominant thought in Strawson, especially not as regards his engagement with the skeptical challenge of determinism.

[78] Strawson, *Skepticism and Naturalism,* 20, 28.

[79] Ibid., 16.

[80] Option (iii) requires some elucidation. The idea here is that framework beliefs are justified because the more specific beliefs they enable us to have seem remarkably reliable/useful/consistent, etc. and that would not be possible unless the framework beliefs themselves were justified.

Arguments based on the supposed inescapability of responsibility-attributing practices figure alongside the compatibilist denial of the generalization strategy discussed in the previous section. The crucial point in both instances is that the truth (or falsity) of determinism is irrelevant to the justification of attributions of responsibility. While the denial of the generalization strategy provided indirect evidence in support of that point by rebutting the incompatibilist who thinks that determinism is relevant because it would function as a standard excuse/exemption, the various arguments from inescapability to be discussed below supply positive reasons why we should stick to the claim that determinism is irrelevant.

Inescapability arguments come in two basic varieties. According to the first variety, determinism is irrelevant to the justifiability of responsibility-attributions because we *cannot* give up the "general framework of attitudes itself".[81] This framework is something we are inexorably committed to no matter how it stands with the thesis of determinism. Construing others and ourselves as responsible agents through our reactive attitudes and feelings is an "essential part of moral life"[82] without which "our existence as social beings"[83] is unimaginable. What kind of incapacity is at issue is elucidated in the 'no justification view' and the argument from naturalism to be discussed below.

By contrast, inescapability-arguments of the second variety appeal not to an incapacity, but to normative considerations. Whether or not determinism could move us to radically alter our perspective or reshape the overall framework of attitudes, for overwhelmingly good reasons we *should* not do so.[84] These reasons are spelled out in the argument from value and the argument from rationality.

Let me take these arguments one by one. Some of Strawson's formulations appear to support explicitly what I want to refer to as the 'no justification view'. On this reading, recognizing the inescapability of the practice of responsibility is tantamount to recognizing that this practice requires no justification: "the existence of the general framework itself neither calls for nor permits an external reaction justification".[85]

One crucial question concerning the justification of any specific belief or judgement is how far we must go back to strike upon beliefs which can be taken to be basic for the purpose of the specific justificatory task at hand, basic in the sense that they themselves need no justification. As already

[81] Strawson, 'Freedom and Resentment,' 23.

[82] Ibid.

[83] Strawson, *Skepticism and Naturalism,* 33.

[84] The two different modalities–can *vs* should–corresponding to the two varieties of inescapability-arguments are conjoined in the original summary presentation of the Inescapability Thesis at the end of Section 4.2.

[85] Strawson, *Skepticism and Naturalism,* 41. We find the same formulation in Strawson, 'Freedom and Resentment,' 23.

indicated, that question can be read as asking 'how far *should* we go back?' but also as asking 'how far *can* we go back?'. The 'no justification view' is concerned with the latter question (so it belongs to the first variety of inescapability-arguments). The answer it gives to this question is that we cannot go back far enough to ask for the justification of the framework itself because there is no platform, no perspective from which to carry out the justification of the framework itself. To use Gary Watson's phrase, there is simply "no more basic belief"[86] to appeal to in order to obtain the justification or rationale for the overall framework of reactive attitudes.[87]

Again, there are two ways of understanding the claim that there is "no more basic belief". It could mean that the demand for justifying the framework itself calls for a perspective that simply does not exist, it is a view from nowhere.[88] That does not appear to be a very promising line since the perspective very plainly does exist. It may be unstable or false (or for psychological reasons unavailable to us as a practical stance), but incompatibilism *per se* is neither meaningless nor incoherent.

Another way of thinking of this claim is to insist that there is "no more basic belief" because the framework itself is non-rationally grounded. That is to say, the practice of responsibility-attribution rests ultimately on a non-rational commitment, it is "more properly an act of the sensitive, than of the cogitative part of our natures".[89]

Looking at the 'no justification view' in this way becomes more convincing when we recall Strawson's emphasis on the decisive role of emotions and feelings in the practice of responsibility-attributions. If that role is indeed as decisive as Strawson makes it out to be and if it is also true that at some basic level the emotions one entertains and the sensitivities one is hard-wired with neither call for nor permit justification, then we may have the answer to the question why the general framework itself cannot be the object of justification. Asking us to justify it is like asking to justify our most basic emotional propensities and reflexes that we are born with.

There are a number of objections to this proposal, however. Most of these are better rehearsed in response to the more forceful naturalistic ar-

[86] Watson, 'Responsibility and the Limits of Evil,' 255.

[87] Several commentators read Strawson in this way or at least *also* in this way. For example, Stern: "The question whether it is rational to give it up [i.e. the commitment to reactive attitudes] cannot even be raised: rational justification takes place within the framework of basic human commitments", see Stern, 'Freedom, Blame, and Moral Community,' 73. But what Stern thinks cannot be done is precisely what all incompatibilists are doing: they are raising the very question concerning the justifiability of that framework. It will not do to simply tell them that they cannot be doing what they are doing.

[88] As arguably, for example, it would be impossible to have an external perspective of an introspected mental state or subjective experience. That is to say, you can have an external perspective of me experiencing it, but you cannot have a view of how it is for me to experience it, that is, of my experience *qua* mine.

[89] Hume, *Treatise* (I.iv.), 183 quoted in G. Strawson, *Freedom and Belief,* 87.

gument (most importantly, that it conflicts with much of what Strawson himself is saying elsewhere and also that it ignores the fact that incompatibilism is rooted in equally natural intuitions about responsibility), so I will only concentrate here only on the way this suggestion conceives of the connection between emotions and justification. The problem, of course, is that it involves taking an extremely simplistic view of emotions which is not borne out either by the findings of the empirical psychology of emotions or our subjective experience of what it is to entertain or feel an emotion.

To mention only a few rather obvious points in support of this objection: Opinions are of course much divided as to the precise relationship between emotions and cognition, feelings and judgement. However, hardly anyone would make a case for the view that emotions are entirely separated from our rational faculties. In addition, we can usually give perfectly coherent answers to why we entertain a certain emotion in a particular case (and even if we cannot, we will still not think the question meaningless). Further, there are also good accounts available of the rationality of emotion-types themselves. Finally and most importantly, emotions seem to be by and large responsive to reasons (and perhaps the converse is true too, but that only strengthens the argument). Again, this is true of emotional reactions in particular cases, but it is also true to say that we can be reasoned out of entertaining certain emotion types or whole ranges of emotions. If that is all true, then the appeal to emotions will not rescue the 'no justification view'.

I now, therefore, come to the *argument from naturalism.* The crucial premise of this argument is that the commitment to reactive attitudes and feelings is a natural fact. It is a deeply ingrained part, a 'given' of human nature.[90] Attributing responsibility to others and ourselves, praising and blaming are "natural expressions of natural responses to what we see people do".[91] Again this argument appeals implicitly to the influential but often insufficiently examined view of emotions as (i) divorced from our rational, deliberative faculties, and as such (ii) constituting the innermost core of human nature, the stuff we are made of.[92] I have already offered some criticisms of this view of emotions above. However, it is crucial to note that, unlike the 'no justification view' just discussed, the argument from naturalism is not intended to demonstrate the non-rational character of our

[90] Strawson, 'Freedom and Resentment,' 18, 23 and Strawson, *Skepticism and Naturalism,* 33, 39.

[91] Wolf, 'The Importance of Free Will,' 389.

[92] That appeal does not always remain implicit. Consider Strawson's diagnosis in *Freedom and Resentment* that both optimistic compatibilists and pessimistic incompatibilists tend to "overintellectualize the facts" by not paying sufficient attention to the "web of human attitudes and feelings" (23) or his hinting that the objective-reactive distinction might be to some extent read as an opposition between "our humanity and our intelligence" (10).

commitment.[93] Thus Strawson explicitly says that what we are naturally committed to is "that whole web or structure of human personal and moral attitudes, feelings, *and judgments*".[94]

The point of the argument from naturalism can be best understood by asking why the reference to this allegedly natural fact is thought to refute the incompatibilist and, even more importantly, how it is thought to impact on the justification of responsibility-attributions? The answer is that what the appeal to this natural fact is supposed to show is that there is a thoroughgoing psychological incapacity rooted in human nature which makes it impossible for us to give up reactive attitudes and the attributions of responsibility these attitudes are tied up with. But if it is impossible to do so, then it is in vain to argue that we should. There is an 'ought' only where there is a 'can', or in Strawson's own words: "there can only be a *lack* where there is a *need.*"[95] We have not chosen our commitment at the first place, nor can we choose to opt out of it. Hence arguments purporting to produce reasons why we should do so are as idle as arguments as to why we should aim to have eternal life. Inescapability justifies.[96]

The essential difficulty with this version of the inescapability argument is that the crucial premise about reactive attitudes and feelings being inescapable natural facts is far from watertight. As Paul Russell points out in his incisive analysis of this argument, the premise equivocates in an unacceptable fashion between the inescapability of particular reactions (reaction-tokens) and the inescapability of general propensities or dispositions (reaction-types).[97] This is a very useful distinction. First, at the token-level, as a matter of empirical fact there do not seem to be such specific cases in which a certain reactive attitude just is inescapable. We might on many occasions find it difficult to withhold reactions, but there appear to be no specific situations inescapably and irremediably triggering certain reactions. Recall what kind of responses are at issue here. By Strawson's own admittance too, not simple reflexes or gut reactions, but complex responses involving and partly depending on beliefs, perceptions and assessments of what others and ourselves do. It is not very plausible to say that these reactions could ever be wholly beyond our control in any given token instance.

But even more importantly, second, consider for the sake of the argument whether it would make any difference if there were indeed such token-

[93] Insofar, I believe, Galen Strawson's labelling of his father's account as a "non-rational commitment theory of freedom" (see G. Strawson, *Freedom and Belief,* 84 and passim) is misleading.

[94] Strawson, *Skepticism and Naturalism,* 39–my italics.

[95] Ibid., 41.

[96] Strawson is of course not alone in describing responsibility-attributing practices as rooted in human nature and in seeking to present 'rootedness' as being in itself a sufficient justificatory consideration. Similar claims are made in, for example, Honoré, *Responsibility and Fault,* 30. There we also find a variety of different inescapability arguments.

[97] Russell, 'Strawson's Way of Naturalizing Responsibility,' 295-8.

cases. Would in such token-cases the mere inescapability of the response justify the response? I think not. In fact, again, so much is admitted by Strawson himself. After all, he makes it quite clear that whenever an excuse or exemption obtains in a particular case it is obligatory to suspend reactive attitudes. But if that is true, then it cannot *also* be true that inescapability is sufficient to justify some reactive attitude in any given token instance. For it cannot simultaneously be the case in one and the same situation that (i) token-inescapability for reactive attitude X (say, resentment) is true, (ii) token-inescapability justifies reactive attitude X, and (iii) an excuse/exemption makes it obligatory to suspend reactive attitude X.

In the article already referred to, Paul Russell also mentions that type-inescapability may be more plausibly defended.[98] Of course, he is also right to add that type-inescapability is neither here nor there as regards the worries of the incompatibilist.[99] The fact that we have the propensity to react in certain ways in certain situations does not *justify* our actual reactions on any given occasion. Granted, even if determinism was demonstrated to be true we may not be able to alter or radically reshape either this propensity or our basic patterns of behaviour, but that does in no way entail that determinism would not furnish us with reasons to reconsider the justifiability of reactions in particular cases.

I turn now to the second variety of inescapability-arguments, those based on normative considerations. The arguments from value and from rationality are often presented conjointly as a single argument. The thought common to both of them is that even if we were in fact capable of giving up reactive attitudes and the attributions of responsibility with which they are tied up, we should not do so. In other words, even if the naturalistic argument as regards the alleged psychological incapacity rooted in human nature did not go through, the truth of determinism all things considered would not constitute a sufficient reason to renounce our commitment to our ongoing practices of reactive attitudes and responsibility.

As these two arguments, despite this common concern, address somewhat different issues I will discuss them separately. In fact, Strawson only hints at the argument from value in some passages of *Freedom and Resentment.* Thus he mentions the ”gains and losses to human life, its enrichment or impoverishment”[100] that would be caused by giving up ordinary reactive attitudes. But since elsewhere too he highlights the ”very great importance that we attach to the attitudes and intentions towards us of other human beings”,[101] Strawson quite clearly believes that abandoning ordinary reactive attitudes would involve a loss of something we hold dear.

[98] Ibid., 302.

[99] Ibid., 297.

[100] Strawson, ’Freedom and Resentment,’ 13.

[101] Ibid., 5.

The argument that renouncing our commitment to reactive attitudes would involve renouncing something valuable is most succinctly formulated in the writings of authors commenting on or inspired by Strawson's work.[102] Again, the argument is at its most persuasive if reactive attitudes are equated with the whole range of attitudes and feelings available to us in interpersonal relationships *and* if the objective attitude is presented as a wholly detached stance towards fellow human beings. We have seen that there are good reasons to steer clear of both of these claims. At the same time, one can still see why Susan Wolf, for example, would insist that "a world without reactive attitudes would be a tragic world of human isolation".[103] Although life without reactive attitudes may not be as bleak as Wolf would have it–in part because the reactive-objective distinction is far less clear-cut or irreconcilable than is implied by Strawsonians–the give-and-takes of resentment, guilt, forgiveness, gratitude, anger, etc. form undoubtedly an important part of our lives enabling us to see, assess and shape human behaviour in ways which would not be open to us without these reactive attitudes being accessible.[104] It may also be true as Bennett says that reactive attitudes actually create interpersonal relations because reactive attitudes function in practice as forms of address to other members of the moral community.[105] If true, that would make us even more entitled to regard reactive attitudes as something valuable.[106]

However, various presentations of the argument from value involve a characteristic ambiguity that is worth discussing here because it greatly influences the success of the argument from value itself. Thus it is not clear whether commitment to reactive attitudes (our 'rootedness' in them) is to be thought of as constituting one significant value among others, or rather it is to be thought of as valuable by virtue of being the pre-condition of all other values in human life. The suggestion that reactive attitudes are important forms of moral address pointing towards or preparing for interpersonal relationships or Wallace's idea that responsibility-entailing reactive

[102] See esp. Bennett, 'Accountability,' 29-30; Wallace, *Responsibility and the Moral Sentiments,* 99-100 and Wolf, 'The Importance of Free Will,' 391-2, 400-2.

[103] Wolf, 'The Importance of Free Will,' 400.

[104] Note that this is a different argument from that made in Sher, *In Praise of Blame,* 123-48 who contends that the value and importance of (at least some) reactive attitudes derives from the fact that morality itself is inseparable from the practice of blaming and praising.

[105] A suggestion also taken up by Watson, see Watson, 'Responsibility and the Limits of Evil,' 267, 269, etc. For Bennett's proposal, see Bennett, 'Accountability,' 42-5.

[106] Note that the argument from the value of reactive attitudes is compatible with criticisms of certain reactive attitudes–the usual suspects are guilt and resentment–as unacceptably cruel, vindictive, parochial or too closely tied up with only one of many possible ethical outlooks on life. Such criticisms (going back at least to Nietzsche) have been recently voiced by Williams in his *Shame and Necessity* and his *Ethics and the Limits of Philosophy,* but also by Baier, 'Moralism and Cruelty: Reflections on Hume and Kant,' and Wertheimer, 'Constraining Condemning' as well as others.

attitudes are tied up with specific expectations to which we hold others and ourselves implies the view that commitment to the web of reactive attitudes embodies commitment to one specific value or value domain, which, albeit important, is not the pre-condition of the possibility of commitment to all other values as a kind of 'super-value'. Wolf, by contrast, construes this particular commitment as the *sine qua non* of all meaning and value in human life when saying, for example, that: "living in accordance with the fact that we are not free and responsible beings [the consequences of accepting the truth of determinism] would require us to give up all our values".[107]

The 'super-value position' attributed here to Wolf is unattractive for rather robust-seeming reasons. First, it presupposes the equivocation of responsibility-attributing reactive attitudes with the whole range of interpersonal attitudes, emotions, etc. which equivocation was already criticized as implausible. But, second, quite apart from that issue, there is a good case to be made that the truth (or falsity) of determinism would leave many of our values–aesthetic, moral and otherwise–quite intact. It is hard to see why in a deterministic world there could not be beautiful objects or even virtuous deeds. In other words, one senses another equivocation on which Wolf's argumentation hinges, namely that between values in general and practical values generating reasons for action in particular.

It could be argued of course that determinism undermines not values in general, but specifically the meaningfulness of practical, action-guiding 'oughts', possibly because such 'oughts' require the ability to do otherwise, an ability perhaps incompatible with determinism.[108] But that of course is a mere reiteration of one kind of incompatibilist argument, we are no longer dealing here with the 'super-value' argument. And this limited argument says nothing about the original claim scrutinized here, namely that the reason why we should not give up reactive attitudes is because we would be losing something supremely valuable, in fact everything that is of value to us. Thus we have not been shown that the practice of reactive attitudes and responsibility-attributions is practically inescapable because the price to be paid for giving up this practice would be prohibitively high, namely the loss of all value and all meaning in our lives. This, I believe, is a significant finding which I will also come back to in the course of discussing the argument from rationality because it also appears to undermine Wolf's other claim that the truth of determinism is irrelevant to the rationality of our practices.[109]

If, on the other hand, the commitment to reactive attitudes does not constitute a super-value, but rather just one value among many others, then the argument from value is no longer very forceful. It articulates an impor-

[107] Wolf, 'The Importance of Free Will,' 401-2.

[108] Although one could object to this that even in a deterministic world there could be valid reasons for action.

[109] Ibid., 404.

tant observation about the practice of reactive attitudes and responsibility-attributions being one significant source of value in human lives (highlighting the importance of which is one of the great merits of the Strawsonian theory of responsibility). But it does not show that this practice is inescapable for any value-oriented human being. Moreover, two points already made at an earlier stage bear repeating here. First, going for the objective attitude is itself a value-guided choice (recall the therapist-patient relationship or the fact that we have normative reasons for absolving the incapacitated from responsibility). Second, incompatibilism itself may be construed as a value-oriented position as well. One way to explain why the incompatibilist is pessimistic about responsibility is precisely that if determinism is true, then it is inappropriate because unfair to attribute responsibility to any agent or for any action.[110]

Finally, the argument from rationality concerns the nature and scope of the reasons the thesis of determinism could furnish us with: "It is a question what it would be *rational* to do if determinism were true, a question about the rational justification of ordinary inter-personal attitudes in general".[111] The idea is that determinism is irrelevant (even if we set aside the argument from naturalism and the issue what we are psychologically capable of) because it does not provide us with reasons that would impact on the justifiability of responsibility-attributions.

This argument is frequently combined with the argument from value. Thus, for example, Wolf maintains that *because* determinism is inimical to the existence of any value it can give us no reason at all which could impact on the justification of responsibility-attributing practices, or for that matter, on the justification of any normative practice.[112] Strawson's position is less radical (unlike Wolf, he does not embrace the argument from super-value), but he too appears to suggest in certain passages of *Freedom and Resentment* that it would not be rational to heed in our practical choices the potential consequences of the thesis of determinism *because* the losses caused by doing so would just be too great.[113]

But the argument from rationality can also be read as making a somewhat different point from the argument from value. The point on this reading is not that determinism does not produce *strong enough* reasons to impact on the justifiability of our reactive attitudes (the commitment to these being too valuable), but rather that determinism does not produce the right *kind* of reasons and hence it cannot impact on the justifiability of reactive attitudes and the attributions of responsibility they are bound up with.

Thus we get yet another sense in which the practice of reactive attitudes and responsibility-attributions could be said to be inescapable. The bounds

[110] On the latter point, see Wallace, *Responsibility and the Moral Sentiments,* 102.

[111] Strawson, 'Freedom and Resentment,' 13.

[112] Wolf, 'The Importance of Free Will,' 386, 403, etc.

[113] See esp. Strawson, 'Freedom and Resentment,' 13.

of the practice are co-extensive with the bounds of human rationality: "it would not necessarily be rational to choose to be more purely rational than we are".[114] The challenge of determinism cannot induce us to be more rational than we are, rooted as we are in the practice of reactive attitudes, because determinism being a "general *theoretical* doctrine" its consequences (whatever these we may speculate to be) must remain wholly irrelevant to our *practical* choices.

The idea here, therefore, is that determinism or any other metaphysical doctrine of agency of only theoretical import must remain irrelevant to the justification of reactive attitudes and the responsibility-attributions they are tied up with because only practical reasons flowing from values and norms can be a source of justification in this domain. Bennett who embraces this argument says for instance: "reactive feelings cannot be made impermissible by any facts, e.g. the fact that men are natural objects".[115]

It certainly seems an exaggeration to say, however, that the truth of determinism cannot generate *any* reason whatsoever. If someone was to prove determinism true that would certainly give us one kind of reason, namely a reason to believe that certain facts about agents and actions obtain. It is hard to see, therefore, why in Wolf's account, for example, only practical reasons are considered (and then ruled out) as potential consequences of determinism. Even if it is true that determinism cannot give us reasons to "live in accordance with the facts"[116] as this would be a practical choice and matter of practical rationality, determinism can very well give us reasons to embrace various facts as being true, which is a theoretical choice and a matter of theoretical rationality.

If that point is conceded, there are two ways to go. First, one could argue in a Kantian vein that the practical point of view is independent from the explanatory perspective and the relevance of determinism is limited to the latter perspective. The idea is that the perspective of practical reason is distinguished from that of theoretical reason, the latter being concerned with the world of sense and the causal relations amongst the entities which inhabit this world. The perspective of practical reason is independent insofar as, unlike the theoretical perspective, it is not concerned with explanation and prediction, but with reasons figuring in agents' deliberations. Because these are reasons for action, they justify rather than causally explain what people choose to do.[117]

[114] Ibid., 13n.

[115] Bennett, 'Accountability,' 29.

[116] Wolf, 'The Importance of Free Will,' 401.

[117] See Korsgaard, 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations,' 204: "For freedom is a concept with a practical employment used in the choice and justification of action, not in explanation or prediction; while causality is a concept of theory, used to explain and predict actions but not to justify them".

On this account, therefore, the perspective of practical reason is discontinuous with the context of natural scientific theorizing, i.e. with the context of causal explanations including causal explanations of action in general. Although it may be true (as Kant thinks for example) that when moving in the latter context all actions are causally determined–"the positive conception of freedom, then, is not to be given a theoretical employment".[118] But that is irrelevant as regards the justifiability of actions and hence the justifiability of ascriptions of responsibility too. The question of justifiability can only be validly posed from the perspective of practical reason.

The problem with "insulating" the practical perspective in this way[119] is that it is not at all clear that the question of causality is irrelevant to the practical perspective. The very least we can say is that how we explain an action in causal terms will be very much relevant to whether the ascribing responsibility for that action is justifiable or not. This is evidenced most clearly by the relevance of excuses (i.e. local responsibility-undermining conditions) to the justifiability of ascriptions of responsibility.[120] Whether the agent 'could have done otherwise' is not merely a consideration relevant to the theoretical perspective in which responsibility is construed with the purpose of explaining the action in causal terms.[121] It is very much relevant to whether it is right to ascribe responsibility to the agent for that action. If the agent could not help doing what he did, the action will not only be mistakenly described, but the agent himself will be wronged.

But if that is true, then it has not been shown that the truth of determinism cannot be relevant to the rational justifiability of responsibility-ascriptions and the concomitant reactive attitudes. In fact, I have already argued at the end of Section 4.5 (p. 102) that the truth of the thesis of determinism can have practical implications. This is because we tend to think that the appropriateness or fairness of responsibility-ascriptions depends (at least in part) on their being true. But unless it can be shown that the thesis of determinism gives us no reason to doubt that our judgements of responsibility converge on truth (because determinism is compatible with responsibility or because determinism is false), we cannot guarantee that our responsibility-ascriptions and the concomitant reactive attitudes will be rationally justifiable.

The other way of re-phrasing the point of the argument from rationality would be that it exposes an irresolvable *conflict* of practical and theoretical

[118] Korsgaard, 'Morality as Freedom,' 174.

[119] On the Kantian insulation strategy, see Wallace, 'Moral Responsibility and the Practical Point of View,' 159-64.

[120] In fact, so much is admitted by Korsgaard herself: "the very idea of an action's being excusable or forgivable or understandable seems to bring together explanatory and justificatory thoughts", Korsgaard, 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations,' 206.

[121] Pace Korsgaard, see esp. ibid., 197-8.

rationality.[122] What those who accept this way of putting the argument from rationality can then be understood as saying is something like this: "Strawson would have managed to show that there were overwhelmingly good rational reasons–*reasons that even outweigh the concern for truth*–for us to distract our own attention from the falsehood of the non-deterministic assumptions that conditions our practices".[123]

If there is such a conflict we would *nolens volens* have to put up with the fact that what we justifiably believe may conflict with what we justifiably ought to do. The consequence of this is that there would also be two kinds or concepts of justification: justification increasing the likelihood of a belief being true, on the one hand, and justification increasing the likelihood of what it is fair or appropriate to do, on the other. The bitter truth may be that we will sometimes find these two forms of justification at loggerheads. If that is the case, then inescapability indeed justifies but it does not justify everything.

Note however, first, that even though this 'tragic' conclusion is derived from Strawson, it would probably not be accepted by Strawson himself. It is anything but reconciliatory. And note also, second, that this conclusion will not lay the pessimists' doubts to rest. In fact, it would encourage them to continue to look for a libertarian or traditional compatibilist solution.

4.7 Conclusion

The findings of this chapter have been critical as well as positive. On one side, I have argued that despite all its brilliance and depth of insight, the Strawsonian expressivistic theory of responsibility fails, or at best, it represents an uneasy and ultimately unstable compromise between cognitivist and non-cognitivist accounts. We have also found good reasons to question the special, Strawsonian brand of reconciliatory compatibilism as an answer to the persistent challenge posed by determinism and 'pessimistic' incompatibilists. Nor was reconciliatory compatibilism found to be superior to traditional compatibilists approaches.

At the same time, on the positive side, Strawson has made an invaluable contribution to our understanding of the concept of responsibility by stressing the link between attributions of responsibility, on the one hand, and reactive attitudes and emotions, on the other. This is a link that any cognitivist theory of moral responsibility must appreciate and account for. Moreover, there appears to be another important insight contained in the

[122] More precisely, a conflict of theoretical rationality and that specific domain of practical rationality that concerns the rationality of what one ought to do. If my criticisms of Wolf are correct, then we need not assume that there must necessarily be a conflict between theoretical beliefs and *all* evaluative beliefs.

[123] Wiggins, 'Towards a Reasonable Libertarianism,' 300–my italics. This is also Ayer's position in 'Free-will and Rationality,' see esp. 12-3.

argument(s) from inescapability and the view they articulate of the connection between inescapability and justifiability. I have argued that grasping this insight leads ultimately to conclusions–including the diagnosis of a potential conflict between practical and theoretical rationality–which are not necessarily in harmony with the original intent of Strawson's theory.

It remains to be seen whether these insights can be accommodated within the framework of a cognitivist theory of responsibility that dispenses both with the naturalism and non-cognitivistic leanings of 'Strawsonianism'–a task I will undertake in the final chapter of this work. Before that, however, I will turn to a different cognitivist approach to responsibility.

Chapter 5

The Ledger View of Moral Responsibility

5.1 Introduction

In examining theories of responsibility I have relied on two basic distinctions so far. The *first* distinction was drawn between 'being responsible' (judgement) and 'holding responsible' (manifest response). A common weakness of the theories discussed until now was that they either ignored this distinction or obliterated it by expressly equating 'being responsible' with 'holding responsible'. This is a serious problem because, as I argued, ascriptions of responsibility are never sufficient to justify overt sanctioning behaviour and in many cases they are not even necessary to justify such behaviour. Further, the failure to appreciate this distinction makes it difficult to account for the special reason-giving force of those normative consequences which are predicated on ascriptions of responsibility (e.g. punishment, guilt) as opposed to which are not (e.g. anger).

The *second* distinction was drawn between emotional response and cognitive content. The Strawsonian theory was criticized for obscuring the point that it is their propositional content, i.e. the belief that 'someone is at fault' or that a 'wrong has been committed', that gives ascriptions of responsibility their normative edge. Insisting on this does not commit us to the view that ascriptive theory is exclusively "to be conceived as a structured array of propositions or judgements".[1] Quite the contrary, it can be readily granted that "patterns of emotional and practical response" constitute an integral part of this domain as vehicles for expressing "a nexus of distinctive sensibilities, cares, and concerns".[2] But, as we have seen, there are reactions which are appropriate responses to only certain kinds of action, namely those for which the agent is responsible.

[1] Wiggins, *Ethics,* 238.

[2] Ibid.

What I will refer to as the Ledger View in the following recommends itself by taking both of the above distinctions into account. Indeed, I am interested in the Ledger View because it promises a *cognitivist* alternative to non-cognitivist–i.e. Strawsonian or emotivist–as well as consequentialist theories of responsibility. It is a conception of responsibility fully developed only by a handful of authors, notably by Joel Feinberg and Michael J. Zimmerman, but it can be detected as an important influence in the works of many other philosophers past and present.

In the course of exploring this alternative it should also become clear what the label 'cognitivist' stands for. The main reason for classifying a theory as cognitivist is its claim that an ascription of responsibility involves first and foremost a certain judgement concerning the agent to whom responsibility is ascribed. But 'cognitivism' also denotes a metaethical commitment. In the present context, it refers to the position according to which the beliefs underlying ascriptions of responsibility are capable of being true or false and therefore we can in principle know that 'A is responsible for Φ-ing'.

So the discussion of the Ledger View can also be helpful because it raises the question in what sense someone's responsibility constitutes an object of knowledge. Is it the case that when I say 'A is responsible for Φ-ing' I lay claim to *knowing* that 'A is responsible for Φ-ing'? As will be seen, the Ledger View regards such ascriptions as purporting to make factual judgements. Indeed, according to pessimistic adherents of the Ledger View the concept of moral responsibility is deeply problematic precisely because it is in principle impossible to say how ascriptions of moral responsibility could ever be true or false. Therefore, they say, we are forced to embrace a skeptical conclusion with regard to the possibility of knowledge about responsibility.

I will argue that the Ledger View does not separate these two senses of what it means to be cognitivist about responsibility and this is why it flirts with skepticism about responsibility. More precisely, the Ledger View erroneously runs together (i) claims about the metaethical status of responsibility-ascriptions, (ii) substantial claims about the conditions under which one is morally responsible, and (iii) claims about how acting wrongly (or rightly) impacts on the agents standing or status. It is the first group of claims that is about the objectivity and knowability of ascriptions of responsibility. The second group is about what being responsible actually consists in, i.e. what must be true about the action for the agent to be blameworthy or praiseworthy for it. That is, under what circumstances must the action be carried out for it to qualify as an entry into the agent's ledger: is it to be voluntary?, can we avoid or incur moral responsibility by luck?, etc. The third set of questions is what gave the Ledger View its name. It is about how and why the ascription of responsibility for an action can impact on our view of the agent himself, i.e. the explanation of why the wrongness of an action justifies our condemnation of the agent. The central idea here is that ascriptions of responsibility impact on one's standing because everyone has a ledger in which one's blameworthy or praiseworthy acts are entered and that this ledger is one's own in a special way because it is non-transferrable.

The different kinds of worries these three groups of claims address are run together in an argument that issues in a skeptical conclusion with regard to ascriptions of responsibility. What proponents of the Ledger View argue is that for judgements of responsibility to be objective there would have to be facts 'out there' (existing, ontologically speaking, independently from and prior to practical reasons and values) to which judgements of responsibility answer. This is because only if such facts guarantee the objectivity of judgements of responsibility can these judgements be regarded as justified and fair (a question which is not only of theoretical interest since many of our typical reactions to other people's behaviour depend on whether there action occasions an entry into their ledger, i.e. whether they are found responsible or not). However, so pessimistic proponents of the Ledger View, the question of the agent's responsibility for his action is forever underdetermined by the facts. In other words, there is no fact of the matter as to whether the conditions which would have to be met for the agent to be responsible for his action–e.g. voluntariness, immunity to luck, etc.–have indeed been met or not. Thus the Ledger View ultimately leads to skepticism about moral responsibility: for judgements of responsibility to be justifiable, they would have to be objective, for them to be objective, there would have to be non-moral facts about responsibility. But there are no such facts. Therefore, moral responsibility is not something that can be justifiably predicated of agents on account of what they have done and ascriptions of moral responsibility are vacuous and have no cognitive meaning.[3]

That conclusion would have wide-ranging consequences for the legitimacy of our responsibility-attributing practices and more broadly for morality as a whole. In this chapter, I will first summarize the main tenets of the Ledger View and show how they can lead to skepticism about moral responsibility. Then I will seek to answer the skeptical challenge. My reply to this challenge will rest on the following two considerations. On the one hand, I will argue that the criteria for what is to count as a 'fact', with which adherents of the Ledger View tacitly operate, are mistaken. For instance, the vagueness of the conditions of responsible agency was one of the reasons why the Ledger View found ascriptions of responsibility to be underdetermined by facts 'out there'. But many simple descriptive statements such as 'this cat is white' are also vague. This, however, does not mean that there is no fact of the matter as regards the cat's whiteness. Nor does it mean

[3] Because it insists that responsibility-ascriptions are to answer to brute facts 'out there' the Ledger View belongs to the family of objectivist theories introduced in Chapter 2, see Section 2.4. Where the Ledger View differs from other members of this family discussed there is its skepticism whether such facts could ever be individuated.

that the descriptive statement could not be unambiguously true/false. Or if there is a problem here it is not specific to ascriptions of responsibility (in fact, arguably, all expressions of a natural language are vague).

On the other hand, I would like to show that thinking of judgements of responsibility as dependent for their appropriateness not on the so-called brute facts of the physical world, but rather on facts such as the existence of practical norms or values, does not entail that "moral responsibility would be undecidable in principle"[4], nor that these judgements would then be dependent on one's subjective inclinations or emotional dispositions. If that is correct, then whatever else is true, the justifiability of ascriptions of responsibility will not depend on our success in finding non-normative facts out there to which these ascriptions are supposed to answer.

5.2 The Ledger View summarized

The Ledger View rests on a principled distinction between judging an agent to be responsible, on the one hand, and responding to his action in some way, on the other. Michael Zimmerman puts this in terms of the "absolutely critical" opposition between appraisability and liability where "an agent is appraisable if he is deserving of a certain type of judgement; an agent is liable if he is deserving of a certain type of treatment".[5] Similarly, Feinberg says that "being 'to blame' and being subject to further blaming performances are two quite distinct things: the former is usually necessary but not always sufficient for the latter".[6] It is worth noting that the distinction is also echoed by authors who are not necessarily adherents of the Ledger View[7] but are critical of available non-cognitivist/consequentialist alternatives. This is because, as we have seen, a common weakness of non-cognitivist and consequentialist accounts of responsibility is that they seem unable to account for our basic intuition that an ascription of moral responsibility may be justified and yet *expressing* this ascription may not be, i.e. that on the whole different sorts of considerations justify ascriptions of responsibility and the imposition of normative consequences for something the agent has done.[8]

But if an ascription of responsibility need not entail any form of overt reaction to what the agent has done, if it is not itself some form of action,

[4] Feinberg, *Doing and Deserving,* 32.

[5] Zimmerman, *An Essay on Moral Responsibility,* 4.

[6] Feinberg, *Doing and Deserving,* 128. See also ibid. 30, 52, 188, etc.

[7] See among others Scanlon, 'The Significance of Choice,' esp. 169: "[...] the origin of this distinctive force [is located] in what is claimed about the person judged" and Wallace, *Responsibility and the Moral Sentiments,* esp. 33: "We need an account of the cognitive dimension in reactive attitudes that will enable us to draw the right kind of line between the moral and the nonmoral reactive attitudes, and between moral reactive attitudes and other kinds of moral sentiment."

[8] For details of the different types of normative consequences, see Section 2.6.

then what is it? The Ledger View takes its name from the distinctive answer it gives to this question, namely that "moral responsibility is liability to charges and credits on some ideal record, liability to credit or blame (in the sense of 'blame' that implies no action)."[9] That is to say, "the doing of the untoward act can be charged to one, or registered for further notice, or 'placed as an entry on one's record'."[10]

Zimmerman repeats this formulation: "blaming someone may be said to constitute judging that there is a 'discredit' or 'debit' in his 'ledger', a 'negative mark' in his 'report-card,' or a 'blemish' or 'stain' on his 'record' [whereas] someone is praiseworthy if there is a 'credit' in his 'ledger'".[11] Such metaphors are adopted by a number of other authors who are more or less sympathetic to the Ledger View. For example Richard Swinburne says that: "Through his past failure the guilty one has acquired a negative status, somewhat like being unclean[...] Both objective and subjective guilt [guilt in which I believe I have done something wrong] are stains on a soul... Such[...] is the common understanding of moral guilt, the status acquired by one who fails in his obligations."[12] Or somewhat less sternly, Jonathan Glover: "[...] involved in our present practice of blaming is a kind of moral accounting, where a person's actions are recorded in an informal balance sheet, with the object of assessing his moral worth".[13]

Note that the 'ledger' or 'record' is thought of as an ideal or as a metaphor. *Actual* ledgers, report-cards, i.e. overtly made (written or verbal) evaluations of one's actions or character are manifest responses. Such overt evaluations can serve a variety of purposes and may be drawn up for quite different reasons, but they are not the ideal ledger at issue here.[14] Note also that according to the Ledger View the ideal record is to be distinguished from evaluations of one's character or overall moral worth as a person.[15] Individual entries in the ledger are occasioned by what the agent

[9] Ibid., 30.

[10] Ibid., 124. See also 188, etc.

[11] Zimmerman, *An Essay on Moral Responsibility,* 38.

[12] Swinburne, 'The Christian Scheme of Salvation,' 15-18.

[13] Glover, *Responsibility,* 64.

[14] On the variety of formal records ("[...] found in offices of employment, schools, banks, and police dossiers... full of grades and averages, marks and points, merits, demerits, debits, charges, credits, and registered instances of 'fault'") used in institutional contexts and their "informal analogue (reputation)", see Feinberg, *Doing and Deserving,* 124-5.

[15] It is clear that individual entries on the ideal record can form the justificatory basis for imposing various normative consequences, e.g. punishment. But it is a difficult question whether the individual entries add up to anything so that any person's record as a whole could form the basis of further evaluative judgements, i.e. ones concerning the agent's overall moral worth (if indeed there is such a thing). On Feinberg's account "there is no rational way of toting them [=the ledger's credits and debits] up and balancing them off apart from our various and divergent practical purposes", ibid., 54. A further point to note is that the record is focused exclusively on what the agent *does* (or intends to do) but looking at one's actions does not say everything about how one *is.* The record may not

does and "a good person may be blameworthy on occasion, and a bad person praiseworthy".[16] So what we are concerned with here is something different from both manifest evaluative responses to action or overall assessments of a person character. The ledger of moral responsibility is an ideal record in which nobody actually enters anything: "A person can be praiseworthy or blameworthy without anyone's being aware of this, without anyone's taking note of it, without anyone's actually praising or blaming him".[17]

So far so good. However, we make judgements of responsibility all the time. But if the agent's responsibility obtains independently from our actual overt reactions to the agent as was claimed above, then how shall we characterize the nature and function of these judgements? The answer, according to the Ledger View, is this: our judgements of moral responsibility, if accurate, *track* the existence of credits or debits in the agent's record which credits or debits are occasioned by what the agent does. So it is concluded that blameworthiness and praiseworthiness are "strictly nonmoral type[s] of worthiness; [they are] a matter of truth or accuracy of judgments".[18]

In short, on this account the agent's responsibility and therefore his blameworthiness or praiseworthiness for an action is a fact that obtains even if no one ever thought it did. As Feinberg puts it: "moral responsibility must be read off the facts or deduced from them".[19] Ascriptions of responsibility on this account amount to *discovering*, rather than *deciding* that the agent is in fact responsible.[20] These ascriptions are "committed totally by the facts".[21]

It also follows from this characterization that questions of moral responsibility leave no room for discretion of the judge, whoever the judge may be.[22] This is because given that they are wholly determined by the facts 'out there' we expect such judgements to be perfectly precise.[23] Once all the facts are in there is nothing left to deliberate: a valid ascription of moral responsibility mirrors a state of affairs 'out there'–"it is true to the

be fully informative about whether the agent is trustworthy, caring, gentle, perceptive, wise or cruel, self-indulgent, insensitive because these qualities are not fully revealed in the agent's actions not even over the long haul. And yet these qualities are relevant to the agent's moral standing. That is to say, on the basis of consulting the record alone (even if epistemic difficulties are set aside for the moment), we will certainly not know everything about that person that may be relevant to our responses to him, no matter how extensive that record may be.

[16] Zimmerman, *An Essay on Moral Responsibility,* 39.

[17] Ibid.

[18] Ibid., 38.

[19] Feinberg, *Doing and Deserving,* 31.

[20] Ibid., 141.

[21] Ibid.

[22] Ibid., 31.

[23] Ibid.

facts"[24]–but it does not add anything to (or take away anything from) what is anyway the case.

A further consequence of this view is said to be that ascriptions of responsibility are to "hold independently of any purposes, goals, or policies",[25] i.e. independently of all extraneous considerations such as, for example, that the making of the ascription may deter this or other agents from similar actions in the future or that it would contribute to the agent's moral improvement or that there would be a "functional need for a decision" to settle the matter of the agent's responsibility.[26] All these considerations can be relevant to the justification of certain reactions to the agent, but they are neither here nor there when it comes to the question whether the agent is in fact morally responsible.

This is why Feinberg can also say that there is an "unqualified finality" to ascriptions of moral responsibility (presumably only if the judgement of responsibility is itself correct, although Feinberg fails to add this).[27] Whatever the agent may do or be done to later on, he remains blameworthy/praiseworthy for that particular action which occasioned the entry into the ideal record. Thus for example that the agent later undergoes much suffering is irrelevant to the question of his blameworthiness even if that suffering is a consequence of the action for which he is ascribed responsibility. Equally irrelevant is his readiness to repent. The entry into the ideal ledger is indelible. According to Zimmerman's formal definition: "if S wills e at T and is culpable [=blameworthy] for this, then he is culpable at T and *forever* thereafter for it" (where S is the agent, e is an event that the agent believes to be wrong and T is a point in time).[28]

These claims make up the core of the Ledger View. To this, Feinberg adds a number of further claims which are perhaps not essential to the Ledger View but well characterize the conception as a whole and are therefore worth rehearsing here.

First, Feinberg frequently repeats that should ascriptions of moral responsibility meet the above requirements, they would be "superior in ra-

[24] Zimmerman, *An Essay on Moral Responsibility,* 38.

[25] Feinberg, *Doing and Deserving,* 31. See also ibid., 41.

[26] The absence of an imperative to decide is, among others, what distinguishes ascriptions of moral responsibility from judgements of legal responsibility on this account. On the normative significance of the imperative to decide in legal cases, see Dworkin, 'Objectivity and Truth: You'd Better Believe It,' 137ff. The expression quoted above is taken from the Dworkin's article.

[27] Feinberg, *Doing and Deserving,* 31.

[28] Zimmerman, *An Essay on Moral Responsibility,* 40. One is reminded here of Russell's declaration in his *History of Western Philosophy* (quoted in Wiggins, *Ethics,* 205.): "I think that particular events are what they are, and do not become different by absorption into a whole. Each act is eternally part of the universe; nothing that happens later can make that act good rather than bad or can confer perfection on the whole of which it is part."

tionality"[29] over ascriptions of legal responsibility made within a particular legal system. Anticipating criticisms to be made in the following section, it should be noted already at this point that it is hard to see why judgements wholly determined by the facts leaving no room for the judge's discretion would represent a higher degree of rationality than judgements which also take into account various other normative considerations. It seems fair to speculate that judgements of moral responsibility meeting the above criteria are assumed to be more rational by Feinberg because their being "committed totally by the facts" is supposed to endow them with greater objectivity as opposed to judgements leaving room for discretion and influenced by all kinds of pragmatic considerations. Such judgements are bound to be less objective, it is suggested, because they are influenced by extraneous factors and consequently do not reliably track 'what is the case'. They are unavoidably subject to change and variation. Without arguing the point against the Ledger View prematurely (why is objectivity equated with 'being determined by the facts'? and why only 'brute' facts are regarded as facts?), it is already worth calling attention to the pattern of thought that comes to the fore here. Superior rationality is associated with judgements which are taken to be objective on account of being wholly determined by facts. The objectivity, and hence rationality, of norm-dependent judgements (such as legal judgements) is assumed to be inferior.

Second, in an attempt to make skepticism about moral responsibility appear less threatening, Feinberg adds that "in contrast to judgments of legal responsibility, which are forced by the circumstances, judgments of moral responsibility can often be safely avoided, for nothing practical need hinge on them".[30] There are a number of points to be noted about this claim.

One thing is that this claim is inconsistent not only with the point made in Chapter 2 that many normative consequences are predicated on ascriptions of responsibility, but also with Feinberg's own views. As Feinberg himself admits, moral responsibility is necessary for punishment.[31] That is already enough to show that much practical hinges on judgements of responsibility. Moreover, as was argued earlier, responsibility is necessary to justify not only punishment but a whole range of normative consequences (among them many of our reactive emotions). The impossibility of making valid judgements of responsibility would seriously question whether anyone is ever entitled to impose such normative consequences. In addition, arguably, the perennial debate about the problem of freewill is also driven to

[29] Feinberg, *Doing and Deserving,* 30. See also ibid., 37, 41.

[30] Ibid., 30, 41.

[31] See his whole essay on 'The Expressive Function of Punishment' in *Doing and Deserving,* esp. 98-9.

a large extent by a recognition of the normative importance of responsibility-attributing practices.[32]

Further, there is a deeper difficulty concerning the place of responsibility-ascriptions in morality as a whole. Suppose it turned out that the concept of moral responsibility was "vacuous"[33] or "inapplicable".[34] How would this skeptical conclusion reflect on the validity and consistency of our moral judgements concerning wrongdoing, duties, character, virtues and vices, etc.? In general, can we have morality without responsibility?

This is a question raised by Bernard Williams which will be discussed in Section 5.4.3 in connection with moral luck below and then more generally in Chapter 6.[35] As will be seen, Williams is particularly concerned with the question whether we can make sense of responsibility without the requirement which in his view is central to morality, namely that moral judgements are to disregard contingent factors of any kind.[36] Reflection on the seeming impossibility of meeting this requirement leads Williams to conclude that if we cannot have morality and responsibility together that could be one reason to give up morality (together with its alleged obsession about luck) and to hold on to a more basic notion of responsibility.[37] As will also be discussed below, Thomas Nagel is more pessimistic about the possibility of salvaging a coherent notion of responsibility. However, he is perhaps more optimistic than Williams about the chances of morality surviving without responsibility or only with a reduced form of responsibility.[38] In any case, the reference to these arguments by Williams and Nagel already indicate that, contrary to Feinberg's perhaps somewhat incautious claim, responsibility-attributions have practical importance also because the lack of a coherent and justifiable notion of responsibility could cast doubt on the coherence and justifiability of moral principles and their requirements as well.

5.3 The Ledger View criticized

In what follows I want to raise various considerations supporting the conclusion that the Ledger View represents an unstable form of cognitivism about responsibility. I would like to show that the main reason for this is a misunderstanding concerning the implications of what it means to construe an ascription of responsibility as a judgement. It is because of this

[32] See G. Strawson, 'The Impossibility of Moral Responsibility,' 8: "It is a matter of historical fact that concern about moral responsibility has been the main motor–indeed the ratio essendi–of discussion of the issue of free will".

[33] Feinberg, *Doing and Deserving,* 36.

[34] Ibid., 32

[35] See Section 6.3, p. 158f.

[36] See Williams, 'Moral Luck,' 22.

[37] See Williams, 'Moral Luck: a Postscript,' esp. 243-4 and Williams, *Ethics and the Limits of Philosophy,* esp. 174-202.

[38] See his 'Moral Luck,' esp. 36-8.

misunderstanding that the Ledger View leads to skepticism about moral responsibility. Skepticism about responsibility can be avoided provided one succeeds in clearing up the Ledger View's misunderstanding in this area. Or so at least I would like to argue.

The crux of the problem is the move from the claim that "a person can be praiseworthy or blameworthy without anyone's being aware of this, without anyone's taking note of it, without anyone's actually praising or blaming him"[39] to the claims that (i) blameworthiness and praiseworthiness constitute "strictly nonmoral type[s] of worthiness"[40] and (ii) that hence moral responsibility "must be read off the facts or deduced from them".[41]

I think that by making this move the Ledger View commits us to an error which consists in ignoring that ascriptions of responsibility are not answerable to 'brute' facts of the physical world. That is to say when we look for reasons to justify judgements of moral responsibility we turn to values or norms rather than simply seek to discover what is the case. For this reason we have to question the assertion that blameworthiness or praiseworthiness constitute "strictly nonmoral type[s] of worthiness". If that is true, then ascriptions of responsibility are more akin to valuations than observation statements. I also believe (but will not argue here) that their being value-dependent does not diminish the objectivity of responsibility-ascriptions. The Ledger View overlooks this because from finding (correctly) that judgements of responsibility are independent of our actual reactions it rushes to the conclusion that the agent's responsibility for an action must constitute a fact independent from any normative perspective. It is hardly surprising that in consequence of this move the Ledger View comes close to or even embraces skepticism about responsibility since its quest for responsibility-facts 'out there' cannot but fail.

Therefore, I now turn to the question why the Ledger View leads to skepticism about moral responsibility. In the following section, I will examine how the Ledger View deals with some of the basic quandaries of moral responsibility in an attempt to show that these difficulties may indeed be insoluble within the conceptual framework of the Ledger View. But the upshot of that discussion is not that we should embrace skepticism about moral responsibility (as pessimistic adherents of the Ledger View suggest) but that we may have to give up the framework itself and look for a more viable understanding of cognitivism about responsibility.

[39] Zimmerman, *An Essay on Moral Responsibility,* 38.
[40] Ibid.
[41] Feinberg, *Doing and Deserving,* 31.

5.4 How the Ledger View leads to skepticism: the quandaries of moral responsibility

5.4.1 Vagueness

The first quandary is by no means restricted to responsibility-ascriptions. The *vagueness* of moral concepts is a pressing problem in other areas of morality as well.[42] To cite a frequently debated example: At a distance of one step from the drowning child, I am near that child, when I am at a distance of ten thousand steps, I am far from him. But there is no morally significant difference between a distance of *n* or *n+1* steps. At what point can I be said to have violated my duty to come to the aid of the child in distress? At what point do I become a Bad Samaritan (morally speaking) as opposed to a remote onlooker for whom, given his distance from the pertaining events, it is permissible not to take action? The difficulty of answering that question is due to the vagueness of the relevant notion of 'proximity'.[43]

What all vague terms have in common is that they (i) have borderline cases, (ii) lack well-defined extensions/have fuzzy boundaries, and (iii) generate so-called sorites paradoxes.[44] In the above example, the predicate 'is near' is such a vague term. The classic example of 'is bald' illustrates the features of vague terms even more graphically: (i) some people *seem* obviously bald and some anything but, and yet there are some people who may be classified as belonging to either groups, (ii) the set of bald individuals includes indisputable specimens but since there is no clear cut-off point for baldness (e.g. bald$\leq$10000hairs?) the number of persons belonging to this group appears to be indeterminate. As for (iii), a person with no hair on his head seems indisputably bald, but so does the fellow with only one hair, two, three, etc. At no point will the adding of another hair perceptibly alter the condition of baldness. Consequently, by iterating the addition frequently enough it would follow that we are to describe a person with 1 million hairs on his head as bald too. All the premisses appear true and yet this conclu-

[42] Feinberg himself discusses the moral significance of vagueness in Feinberg, 'The Moral and Legal Responsibility of the Bad Samaritan,' esp. 189-94.

[43] Does this example raise a genuine problem at all? Some complain about the very attempt to model moral requirements "on a view of the world in which every happening and every person is at the same distance" (Williams, Moral Luck, 37). Only on such a view, so the complaint, is it meaningful at all to think that an exact boundary must be found. Conversely too, the impossibility of finding such a boundary can be argued to undermine that view itself. In short, the sorites-type paradox (see below) that the search for such a boundary generates amounts in effect to a *reductio* of the understanding of moral requirements as applying irrespective of one's location in time and space. This paradox is thus interpreted as yet another expression of the "genuine pathology of moral life" (Ibid., 38) or at least as another "fault line of morality" (Williams, 'Postscript,' 240).

[44] The following discussion of vagueness relies on Keefe and Smith, *Vagueness.*

sion appears to be equally obviously false. The same kind of paradox can be generated for all vague terms.[45]

These features are all based on the fact that vague terms appear to involve genuine indeterminateness in the sense that "no amount of information can decide their applicability".[46] Nor is vagueness restricted to terms used in certain specialized contexts and not even restricted to normative terms in general. Bertrand Russell argues that vagueness is ubiquitous in natural languages[47] and Raz goes as far as saying that "all, and not only some, nouns, verbs, adverbs, and adjectives of a natural language are vague".[48]

So does vagueness present a special problem for ascriptions of moral responsibility in particular? Feinberg appears to think it does. Vagueness is omnipresent in responsibility-ascriptions. We are confronted with what is to all appearances ineliminable vagueness when determining what we are to be held responsible for, when examining the agent's intentions and motivations, when considering whether a responsibility-undermining condition (an excuse or exemption) is applicable or not, and so on.

For instance, it is commonly accepted that insanity ought to exempt the agent from responsibility for his action. But it seems that whatever criterion of insanity is applied our standard will be to some extent vague. Thus it is often taken to be a decisive mark of insanity that the agent lacked the capacity to understand that what he was doing was wrong and hence was incapable of forming a guilty mind. However, for the assessment of any capacity we will have to take recourse to some standard of normalcy and hence will be confronted with vagueness.[49] It may be thought that the reason why vagueness crops up in these examples is because they invoke strongly evaluative notions (e.g. wrongness, harm, capacity, subnormality, etc.) But the Bad Samaritan example discussed above is sufficient to show that even such a seemingly innocent term as 'proximity' can generate moral disputes due to its inherent vagueness.

[45] Note that, although this distinction is not always drawn explicitly (see for example Hart, *The Concept of Law,* 124-36), vagueness is not the same as 'open-texture'. The latter is a special kind of indeterminacy that has to do specifically with the generality of sortal terms. For example, we may have trouble in deciding whether the general term 'vehicle' is to apply to roller skates in a particular context (see ibid., 126). This is because that term refers to a class of things but the boundaries of that class are unavoidably fuzzy and there will be borderline cases. Insofar 'vehicle' displays features (i) and (ii) of vague terms. But because 'vehicle' is not a scalar concept, it does not generate sorites, so feature (iii) will be lacking. Conversely, it can be argued that the predicate 'is bald' though vague is not open-textured. In any case, both open-texture and vagueness create similar kinds of problems in normative contexts.

[46] Keefe and Smith, *Vagueness,* 2.

[47] See ibid., 4-5.

[48] Raz, 'Legal Reasons, Sources, and Gaps,' 73.

[49] For example: '*A* frequently Φ-ed and therefore can be judged to be suffering from severe mental illness.' But how many occasions of Φ-ing should be taken to be severely subnormal? Frequency is as vague a term as any.

In general, there are different ways of dealing with vagueness: sometimes vagueness can safely be ignored, dealt with by employing comparisons ('A may not be completely bald but he certainly has much less hair than B') or, most commonly, overcome by just drawing the line somewhere, typically by means of quantifying the relevant limit (e.g. height restrictions for recruitment in the army (say) will state not that 'only tall people can join' but rather that 'any applicant must be taller than x cm'). However, the difficulty according to Feinberg is that when it comes to responsibility-ascriptions such cut-off points may not be defensible from a moral point of view because they are arbitrarily imposed on the basis of considerations extraneous to the merits of the individual case. The idea is that while the kind of pragmatic solutions mentioned here can be relied on whenever there is a "functional need for a decision",[50] these pragmatic solutions will be irrelevant to determining whether X *really* is morally responsible for Φ-ing.

As we have seen, on the Ledger View, moral responsibility is a credit or debit in one's ideal ledger occasioned by what the agent's has done. So whether there is an entry should be determined entirely by what the agent has done. As Feinberg says, questions of moral responsibility ought to be "perfectly precise" and "in no way forced by practical considerations".[51] But then the skeptical conclusion seems to follow almost unavoidably as long as these criteria are insisted upon. Since "we are not allowed to appeal to purposes and policies" to settle indeterminacies caused by vagueness but making valid responsibility-ascriptions would require settling these indeterminacies, it appears to follow that moral responsibility is "absolutely undecidable in principle and therefore inapplicable".[52]

The argument that vagueness presents a specific problem to judgements of moral responsibility permits two different readings. On the *first* reading, the argument is a *meta-ethical* one. It contends that moral responsibility is absolutely undecidable because due to the ineliminable vagueness of terms used in responsibility-ascriptions it is a mistake to attribute truth-values to such ascriptions. If that is the intended reading, the argument from vagueness involves a *non sequitur.* None of the main theories of vagueness assume that the problem of vagueness would imply wholesale skepticism about truth. Thus we have 'conservative' solutions which retain classical two-valued logic and semantics. According to these, vague terms do have well-defined extensions and predications about borderline cases are either distinctly true or distinctly false. Our inability to see that indicates merely an *epistemic* limitation, not the inconsistency or incompleteness of two-valued logic. Non-classical theories (e.g. supervaluationism) do concede that the truth-value of borderline predications is indeterminate or strange or that they do not

[50] See p. 123 above.

[51] Feinberg, *Doing and Deserving,* 30.

[52] Ibid., 32.

have a truth-value and that as a result many-valued logic may be a more feasible option. However, even these theories allow for straightforwardly true or false predications.[53] The same applies to moral judgements: while there is always a "penumbra of debatable cases" the penumbra can by definition only appear around a "core of settled meaning".[54] In short, it does not follow from the ubiquity of vagueness that ascriptions of moral responsibility could not be true or false.

The upshot is, *first,* that vagueness is not specific to responsibility-ascriptions and not even to moral judgements.

Second, it is a mistake to assume that the presence of vagueness would exclude access to a robust notion of truth. To see this, note a) that vague predicates admit of degrees (this is precisely why they generate sorites) and hence of comparisons too. The proposition 'X is tall' may not admit of a determinate truth-value but the proposition 'X is taller than Y' does (this is why the conclusion of sorites is unacceptable). Note also b) that instantiations of a predicate can be said to be vague only on the assumption that there are determinate instantiations of the predicate too. For instance, whether 'X is bald' will be vague, only on the assumption that there is at least one unambiguously bald person and one unambiguously hairy person. Otherwise we could not say relative to what X's baldness is supposed to be vague. Accordingly, even non-classical theories of vagueness, while they deny bivalence, do not deny that some applications of vague predicates are determinately true or false. And note finally c) that the fact that a predicate is vague does not tell us anything about the determinateness of the piece of reality to which a given application of the predicate refers. That is to say, it may be indeterminate whether the predicate 'is tall' applies to X or not. But that does not show that X does not have a determinate height. The fact that the predicate 'is red' is vague does not show that the surface I can see over there is of no particular colour.[55] In sum, the first reading

[53] Note that I am not proposing here the solution to vagueness put forward by Feinberg (in other writings) and by Dworkin, see Feinberg, 'The Moral and Legal Responsibility of the Bad Samaritan' and Dworkin, 'No Right Answer?,' 68. *That* solution involves dividing cases into distinctly false, distinctly true applications and distinctly vague applications in between. This solution is not acceptable because, as Raz notes, there may be borderline cases between straightforwardly true predications and vague predications, on the one hand, as well as between straightforwardly false predications and vague predications, on the other. In other words, there is indeterminacy about borderline cases themselves, i.e. there may be borderline cases between 'straightforwardly true X is bald' and 'vaguely true/false that X is bald'. In short, vagueness is 'continuous', see Raz, 'Legal Gaps,' 73. But note that even if vagueness is continuous, there can be determinate truth-values. This is because 'being on the borderline' is a relational property of predications located between true and false predications.

[54] Hart, 'Positivism and the Separation of Law and Morals,' 63.

[55] Nor is any help to be had from theories maintaining the vagueness of objects (rather than of predicates). Such accounts pose a general skeptical challenge to our ordinary use of descriptive predicates, not one specific to ascriptions of responsibility. In other

of the argument from vagueness cannot demonstrate that judgements of moral responsibility could not be determinately true or false due to their vagueness–not even if we were to accept the Ledger View's claim that moral responsibility is to be "read off the facts".

On the *second* reading, the argument from vagueness rests on *first-order ethical* considerations. According to this understanding ascriptions of responsibility are unfair or inappropriate because of their ineliminable vagueness. Consider 'contiguous cases' in which everything is held fixed except one parameter. If that difference lacks moral import, then the agent who is evaluated less favourably only on the basis of *that* parameter may be justified in complaining for having been judged unfairly. Now, assume that the differential parameter is instantiated by a vague predicate. For instance, to return to the Bad Samaritan example, imagine that A was 300 steps removed from the drowning child while B was 301 steps away. The difference appears to lack any moral significance. Now if we judge A blameworthy for not rushing to rescue the child but not B, then A may rightly complain that a difference of one step ought not to make any difference to (the degree of) their respective responsibility. But if one step does not make a difference, how many steps will? And what is it about that n number of steps that makes a moral difference? After all, one can always imagine a contiguous case involving n as opposed to $n+1$ steps! Generally speaking, it appears that any rule drawing the line somewhere in an attempt to eliminate vagueness, no matter how reasonably placed that line may be, may always be contested as morally irrelevant and hence arbitrary.

The skeptical conclusion is thus generated by a dilemma on this reading. On the one hand, without drawing lines for the applicability of vague terms the question of moral responsibility will not be answerable. On the other hand, the drawing of any such line appears to be arbitrary from a moral point of view. Therefore, those judged unfavourably on the basis of such an arbitrary rule will be justified in complaining for being appraised on the basis of an *ad hoc* criterion.

The existence of hard cases–where judgements of responsibility are presumptive or tenuous to an extent that they have to be suspended–should not be doubted. Thus contiguous cases trigger our intuition that a small difference in a vague parameter is insufficient to justify divergent moral judgements.

My response to this is that there is no uncontestable, waterproof way of dealing with such cases of vagueness. But I believe that the ubiquity of vagueness is to be taken as evidence of the dependence of responsibility-ascriptions on norm and values, not as a reason to doubt their appropri-

words, here we are faced with a comprehensive skeptical challenge about the determinacy of the truth-values of descriptive predicates which is not restricted to a specific normative context.

ateness or fairness as such. Those 'on the wrong side of the fence' in a contiguous case protest not that there is no fact of the matter with regard to questions of moral responsibility but rather that they have been treated unfairly since the difference to their counterpart was so minute. In other words appeal is made to norms or values and not facts 'out there'. Perhaps contiguous cases require a special treatment. Thus in genuinely contiguous cases we may have to rely on additional norms or values as 'tiebreakers'.

Therefore, I believe that the argument from vagueness will only engender skeptical doubts as long as we insist that responsibility is to be read off the facts. Only then will we have no answer when judgements of responsibility are contested. The discussion of the problem of vagueness gives a first indication in what sense the Ledger View can be said to pose excessive demands derived from a misunderstanding about the nature of the justification of normative claims and a concomitant misunderstanding about the facts which such judgements are supposed to track. I hope to lend more support to this criticism in the following two sections as well.

5.4.2 Causal responsibility

The notions of cause and causation take centre stage with regard to two aspects of moral responsibility: antecedent conditions of action and outcomes. The first aspect will be briefly touched upon in connection with the argument from control in the following section but will not be dealt with here. The second aspect concerns the question to what extent the agent's causal contribution to bringing about an event is relevant to his moral responsibility for the occurrence of that event. The reason why it is worth investigating this latter aspect is that it is yet another "essential characteristic"[56] of moral responsibility the seeming evasiveness of which can engender skepticism about the concept of moral responsibility itself.

The agent's causal contribution is essential because, as will be discussed below, ascriptions of moral responsibility appear to be predicated, among others, upon the agent's having made a causal contribution. But a closer look at this characteristic raises the spectre of skepticism once again because of what seems to be the inescapable dependence of causal ascriptions on one's perspective, conceptual scheme, interests or purposes. That need not be a problem for one's theory of causation because it leaves it open for one to argue that causation is normative.[57] But it can cause headaches for the Ledger View because if indeed there is such dependence, then not even causal responsibility is a matter of 'brute' facts about the relation of constituent parts of the world 'out there'. *A fortiori,* moral responsibility can be even less of a matter of 'brute' facts 'out there'. And that, again,

[56] Feinberg, *Doing and Deserving,* 36.

[57] For the claim that causation is normative, see for example, Thomson, 'Causation: Omissions.'

appears to suggest that ascriptions of moral responsibility cannot be ”read off the facts” and are therefore vacuous and are therefore unjustifiable.

Consider the following quote, for example: ”[. . .]when we assign moral responsibility[. . .] what we mean in morality is to name a causal relation that is natural”.[58] The implication seems to be that, if anything, its relation to causation provides us with an objective handle on responsibility. Of course, it is itself a moot issue what causation is. But the hope that is revealed in the quote is that if there is such a thing as cause and effect, the relationship between them will be an objective matter, not dependent, that is, on one’s interests or purposes. Now, if even this hope turns out to be an illusion, then that may not necessarily be a problem for understanding causation, but it may indeed leave no option but to embrace skepticism about moral responsibility.

Clearly, ascriptions of responsibility for outcomes involve reference to the agent’s causal contribution: ”to be responsible for an outcome, a person or a thing must play some role in causing or failing to prevent that outcome”.[59] In other words, some form of causal responsibility is a necessary condition of moral responsibility. However, there is a good case to be made that causal responsibility, though always necessary, is never sufficient for moral responsibility. It follows that the principal difficulty here is to spell out *what kind* of causal contribution is necessary for the agent’s moral responsibility. Part of this problem is to say how causal responsibility is related to whatever else is required for the agent’s moral responsibility *over and above* his causal responsibility.[60]

[58] Moore, ’Causation and Responsibility,’ 4.

[59] Sher, *In Praise of Blame,* and Feinberg, *Doing and Deserving,* 32.

[60] We are clearly morally responsible for some of our omissions, at least when they are intentional (e.g. I deliberately refrain from wading into the water to rescue a drowning child). Therefore, if causal responsibility is necessary for moral responsibility, then we must accept that there is causation by omission. This is the mainstream view which I also go along with here (but cf. for example Sartorio, ’How To Be Responsible For Something Without Causing It’ for an attempted rebuttal of the mainstream view). Granted, the mainstream view is not without its weaknesses. The *first* difficulty is to provide a description of agency that can accommodate omissions as causes. What exactly causes my *not* wading into the water to rescue the drowning child? For instance, if we go for the theory of agency that actions are bodily movements or states appropriately caused by certain mental events or states of the agent (e.g. belief and desire pairs, intentions, etc.), then we would have to say that some mental event of mine caused my *not* wading into the water. The lesser problem is that if we are to accommodate omissions, then on this theory we would have to say that the lack of movement can also qualify as an action. The other more pressing difficulty is that we would have to identify the mental event that caused my *not* wading into the water. In accordance with the theory that actions are bodily (non-)movements caused by intention (or some other mental state), this mental event must be the forming of the intention (or some other mental state) not to wade into the water. The problem is only that it may be more plausible to say that my *omitting* to form the intention to wade into the water was the cause of my not wading into the water rather than my intention not to wade into the water. But if the intention is irrelevant

Why is causal responsibility necessary? Couldn't the faultiness of the action suffice or even the faultiness of the agent in some respect (e.g. in some character trait) suffice? The problem is that a fault attributable to the agent does in itself not suffice for moral responsibility for outcomes, even if what is faulty was something the agent has done.[61] For example, speeding is faulty. An accident is a harmful outcome. But the speeding driver is *morally* responsible (legal responsibility is another matter) for the accident only if his speeding caused the accident (and not, say, the drunkenness of the other driver involved in the accident or a technical malfunction of the other car). So the causal relationship between the faultiness of the action and the harm caused must be intrinsic, i.e. the harm had to happen because the agent did wrong for the agent to be morally responsible for that harm. In the above example, speeding does not make the driver morally responsible as long as the accident would have happened even if the first driver would not have been speeding.[62]

because what has causal power is the omission of the intention, then we are left looking for a mental event which *is* the causally-efficacious omission with causal power. (Note that merely saying that the omission itself was intentional threatens with an infinite regress.) This argument is spelled out in considerable detail in Sartorio, 'Omissions and Causalism.'

The *second* difficulty is that once it is admitted that omissions can be causes we get too many causes of the same outcome. My omitting to return the library's *only copy* of *Theory of Justice* was the cause of the book not being available for the next session of the Rawls reading group. But Parfit's omission to return the only copy of *Theory of Justice* to the library was the cause of the book not being available too, since had he returned the book, it would have been available. But it would be strange to say that Parfit's omission was really the cause of the book's not being available. I believe, however, that this second difficulty of there being too many omissions poses no *special* difficulty for the claim that causal responsibility is necessary for moral responsibility because it is an instance of a general problem that arises for ascriptions of causal responsibility, namely that because there are too many causes, identifying the relevant cause seems to be dependent on the given context of the causal inquiry. That problem will be discussed below.

[61] In the remaining part of this section I will only be considering the question of responsibility for harmful outcomes. It is, however, tacitly assumed throughout that responsibility for positive outcomes is symmetrical at least as regards the condition of causal responsibility. If that is true the arguments made in this section about moral responsibility for harm apply to moral responsibility for doing good as well.

[62] One special difficulty which arises out of making causal responsibility a necessary condition of moral responsibility is responsibility for making decisions, forming intentions and entertaining desires or beliefs. Are these also to be counted as events which agents can causally contribute to bringing about? It is clearly not true that these propositional states are things in regard to which the agent possesses no causally efficacious powers (as opposed to, say, his being tall). The agent can alter them, restrain them, reflect upon them, etc. It may be true that whatever causally efficacious powers the agent possesses, these are ultimately not sufficient for his moral responsibility (as claimed by the argument from control to be discussed in the following section) but this is a different problem. The concern here was whether the agent could be said to be causally active with regard to his own propositional states if he has causally efficacious powers at all. I think the answer to this question should be in the affirmative.

But why is causal responsibility not sufficient? First, because if causal responsibility were sufficient we could hold bridges morally responsible for collapsing and flower pots for falling on our heads. But even though, as Feinberg says, it is perfectly normal to "use the language of responsibility"[63] when identifying objects or natural events as causes (e.g. 'the heat wave was responsible for the rise in the number of accidents'), objects or natural events can hardly figure as the proper targets of blame and praise or as the objects of the normative consequences predicated on ascriptions of responsibility.

Causal responsibility will not be sufficient for moral responsibility even if we restrict our attention to human agency. Consider the much-cited case of the lorry-driver who ran over a child through no fault of his own.[64] The dispute here is *not* about the claim that lorry-driver's causal responsibility for the child's death is not sufficient for his moral responsibility for the child's death and *a fortiori* not sufficient for those normative consequences which require moral responsibility (e.g. punishment, guilt). Most, possibly everyone, would agree that that claim is highly plausible. The dispute is about the question whether causal responsibility is sufficient for certain responses, not requiring moral responsibility, to be justified on the part of the lorry-driver himself or that of others. If yes, that would imply that we justifiably react differently to agents than to things even if agents could not help causing what they did.[65]

[63] Ibid., 130.

[64] Usually cited from Williams, 'Moral Luck,' 28. The case was first discussed by Dodds in his 'On Misunderstanding the Oedipus Rex.' See Wiggins, *Ethics,* 251n25 for comments on Dodds.

[65] Thus Dodds and Williams call our attention to a peculiar and significant feature of the case, namely the *perspectival asymmetry* it involves. Despite the driver's not being morally responsible we would find it morally objectionable for him to be entirely insensitive to the fact that (though faultless) his causal contribution was crucial to bringing about the harm. His causal responsibility establishes a special perspective and given this perspective we expect him to feel differently and also to react differently than a mere spectator. There are different ways to account for this asymmetry: one could plausibly say that the difference is essentially epistemic, the lorry-driver keeps worrying (and is *expected to* keep worrying) whether there was really nothing he could have done differently to prevent the accident. Given the enormity of what happened he cannot lay his doubts to rest. Also, one could say, as Williams does, that the difference lies in the fact that although the harm was caused involuntarily but it was nevertheless the outcome of an intentional action. Finally, it can be argued that the lorry-driver incurs special obligations because he, at one point (probably when obtaining his driving license) voluntarily *opted into* becoming an active participant in the practice of driving vehicles, a practice which involves well-known and considerably heightened risks of harm to others (also to those who have not opted into this practice). Without wanting to discuss the merits of these alternative explanations, it should be noted that they share the thought that being an agent–that is, having the capacity to make causally efficacious contributions to how the world is and will be–is in itself a significant fact to be reckoned with in our evaluative judgements even if the agent did not bring about an outcome voluntarily and hence is found not to be morally responsible for what he has done. Perhaps this is not entirely true of the last, option-based explanation, which traces back the perspectival asymmetry to the agent's earlier fully

Further support for the claim that causal responsibility is not sufficient comes from cases in which the agent has a legitimate excuse, e.g. non-culpable ignorance of the harmful consequence of his action, that absolve him from responsibility for it. In such cases the harm is traceable to the agent as its cause but the agent cannot be faulted for the harmful consequence.

The threat of skepticism arises, however, when having to spell out *what kind* of causal contribution is necessary for the agent's moral responsibility and how causal responsibility is related to other necessary conditions of moral responsibility. The problem is that the agent must not only be *a* cause of the outcome in question, but also that he must be *the* cause in a special sense. But identifying the cause in this sense in order to "pin the blame" runs into major difficulties.

One worry is that there is an infinite number of factors standing in causal connection with any event so it will be in principle impossible to provide a complete list of the conditions "severally necessary and jointly sufficient for its occurrence".[66] That does not render causal ascriptions arbitrary but it seems that any selection will be inevitably relative to the given context as well as relative to the interests and purposes in the service of which the ascription is made.[67] Therefore any causal ascription is bound to be discretionary, context-relative and presumptive.

So it seems, again, that causal ascriptions call for a decision rather than a discovery.[68] This would mean that when causal ascriptions are contested the basis for the challenge will be not that they misdescribe the case but rather that they present "the less important as the more important".[69]

voluntary choice and hence need not attribute moral significance to causal involvement in itself. The problem with that explanation is that the same perspectival asymmetry obtains in the case of a flower pot which through no fault of mine falls on child's head and kills him. Here there is no question of there being a practice I opted into. And yet, if it is my window the pot fell from my perspective is bound to be different even if I had taken due care to fasten the pot, etc. But in this case there is no rule-governed social institution I chose to become part of. I am merely a participant in ordinary human interactions. If, on the other hand, we widen the notion of rule-governed social institution to include all human interactions and argue that we are duty-bound to bear the costs of the inherent risks involved in such interactions (for example on the grounds that we also benefit from these interactions), then the difficulty may be that we lose grip on the independent normative significance of voluntariness altogether.

[66] Feinberg, *Doing and Deserving,* 142.

[67] See Honoré, *Responsibility and Fault,* 3: "[...] what counts as the cause of an event depends on the purpose of the inquiry". See also Feinberg, *Doing and Deserving,* 142, 148, 202, etc.

[68] Feinberg, *Doing and Deserving,* 141.

[69] Ibid., 146. Because of this, causal ascriptions seem to resemble defeasible applications of legal concepts. Discussing the specific example of contracts but making a general point, Hart says the following about applications of defeasible concepts: "[...] the judge is literally deciding that on the facts before him a contract does or does not exist, and to do this is neither to describe the facts nor to make inductive or deductive inferences from the statement of facts, what he does may be either a *right* or *wrong* decision or a *good* or

But if causal responsibility cannot be defined in terms of necessary and sufficient conditions it will be found that there will be no fact of the matter to determine what sort of causal responsibility is to be deemed relevant. That implies that truth/falsity will not be predicable of ascriptions of causal responsibility and *a fortiori* (since causal responsibility is necessary for moral responsibility) they will not be predicable of moral responsibility either.

At this stage one may object that this characteristic of causal attributions should not present an unsurmountable difficulty for although ascriptions may indeed be context-dependent that is not a problem because the context is fixed in which these ascriptions are made. Thus, for instance, when we want to "pin the blame" for harm we must look for the faulty action as being "the cause" of harm. The point is, in other words, that there may indeed be an infinite number of causal factors contributing to bringing about the event in question but only that (or those) will be relevant which are morally objectionable or substandard.

It will be recalled that there was no such causal factor in the faultless lorry-driver's case. But now suppose the lorry-driver had been found to exceed the speed-limit. Other things being equal, the speeding is clearly the most conspicuous candidate to be identified as the cause of the accident and the ground of the lorry-driver's blameworthiness. Now the problem seems to be, however, that the identification itself of morally objectionable or substandard causal contributions on the agent's part seems impossible to carry out "without recourse to purposes and policies".[70]

Even the assumption that there was only one morally objectionable causal factor leading to the harmful event will not help. *A* hits *B* and as a result *B* is seriously hurt. That is wrong. But what exactly is *A* morally responsible (and blameworthy) for? Imagine that a certain time after the injury *B* passes away. Imagine too that the injury *B* sustained as a result of his encounter with *A* can be shown to have unquestionably contributed to his death. What criteria should decide whether A is morally responsible for *B*'s death by virtue of his causal contribution to it? Proximity in time, for example, will not do. In criminal law the 'year and a day rule' has until recently been frequently applied to homicide cases stating that *A* was to be held legally responsible for murdering *B* if *B*'s death occurred less than a year and one day since *A* delivered the blow. But that rule seems carry no independent moral significance. Consequently, it seems irrelevant to *A*'s *moral* responsibility for *B*'s death. In sum, *A*'s causal contribution was no doubt morally objectionable but was it sufficiently serious to render him blameworthy for *killing B*?

bad judgement and can be either *affirmed* or *reversed* and (where he has no jurisdiction to decide the question) may be *quashed* or *discharged.* What cannot be said of it is that it is either *true* or *false,* logically necessary or absurd." See Hart, 'The Ascription of Responsibility and Rights,' 182.

[70] Feinberg, *Doing and Deserving,* 33.

As with vagueness, the difficulty lies in the fact that no amount of information will help to answer such questions. According to the Ledger View the question of moral responsibility ought to be resolved in this case as in all others by "reading the answer off the facts". But it seems that (as Hart puts it): "facts are dumb".[71] And note also that the arbitrariness of appealing to *ad hoc* rules seems even more objectionable to settle the agent's causal responsibility than to disambiguate vagueness since the problem here is not caused by an inherent feature of natural languages. Rather, the challenge is quite specific to ascriptive theory. It is posed by the impossibility of defining consistent criteria for a necessary condition the meeting of which would be required for making valid ascriptions of moral responsibility.

So once again the topic of causal responsibility confronts us with a dilemma: On the one hand, there appears to be no "natural stopping place" for attributions of moral responsibility. For the agent to be morally responsible he must be causally responsible but we are unable to specify objective, non-normative criteria for the agent's causal responsibility. On the other, importing extraneous rules or making appeals to policy to settle questions of moral responsibility seems irrelevant and therefore morally objectionable.

There are two ways to go from here. We can explore ways to deny that "what [cause] means in morality is to name a relation that is natural".[72] This would be to abandon the search for 'brute' responsibility-facts to guarantee the justifiability of responsibility-ascriptions. I will explore this suggestion in Chapter 6. Alternatively, it can be argued that in light of the difficulties concerning ascriptions of responsibility for outcomes "moral responsibility for external harms makes no sense".[73] That second option does not automatically entail skepticism as it is still possible that agents are morally responsible for their intentions and their related mental states (e.g. volitions), though not for outcomes. But as the next section should make clear that move, due to the ubiquity of luck, will not block skeptical worries, not at least for adherents of the Ledger View.

5.4.3 Moral luck

The idea that luck cannot make a moral difference and should therefore be consistently factored out from our moral judgements is traditionally attributed to Kant. This is because of Kant's emphasis on the irrelevance to the agent's moral responsibility of both contingent aspects of one's character and the actual outcome of what one intended to do.[74] Feinberg raised the

[71] Hart, 'Positivism and the Separation of Law and Morals,' 63.

[72] Moore, 'Causation and Responsibility,' 5.

[73] Feinberg, *Doing and Deserving,* 33.

[74] Although there is some dispute whether Kant actually held this view, see esp. Latus, 'Moral Luck,' n5. Note that the issue of luck is broader than that of moral responsibility. Thus for example it can be debated whether luck should make a difference to assessments of character quite independently from the matter of moral responsibility.

problem of moral luck in contemporary moral philosophy once again, even before the influential pair of articles by Williams and Nagel.[75]

Feinberg contends that it is among the "essential characteristics"[76] of moral responsibility that it "cannot be a matter of luck". That is, "[...] [moral responsibility] must be something one can neither escape by good luck nor tumble into through bad luck".[77] He then goes on to say that the impossibility of meeting this requirement (I will call it the immunity-requirement) leads to skepticism about moral responsibility. In fact, it seems to be his view that the impossibility of meeting this requirement would *in itself* be sufficient to render "the precise determinability of moral responsibility[...] an illusion," making cases of moral responsibility "undecidable in principle".[78] This would be so even if other necessary conditions for justifying ascriptions of moral responsibility could be fulfilled.

The immunity-requirement lends support to the previously made distinction between judgements of responsibility and manifest responses to action. Whatever we think of the validity of the immunity-requirement for moral responsibility, it is plausible to argue that the immunity-requirement applies differently to how we judge the agent's responsibility as opposed to how we think he ought to be treated in consequence of his action. In fact, it seems that immunity to luck is not required for the justification of a number of different responses to action. Even though we think that there is no difference in terms of blameworthiness, *mens rea* or moral fault between Adam and Bill who slapped two people in one and the same brawl, in one and the same pub, we may have good reasons to respond to Adam more harshly who, unlike Bill, through his bad luck happened to slap a hemophiliac who died as a consequence of the slap.[79]

Generally speaking, we have good reasons to make different overt responses to actions involving an equal degree of blameworthiness but leading to different outcomes, even if the difference in outcomes is due to contingent factors. These reasons can include policies of deterrence or, for instance, the difficulty of proving the agent's responsibility so that actual outcomes of action must be taken into account if any judgement at all is to be made. But of course even if it is true that the immunity-requirement is not presupposed by the justification of some forms of sanctioning behaviour, that is irrelevant as regards the claim that judgements of moral responsibility themselves must be immune to luck and why that should be so.

Feinberg's various examples already foreshadow Nagel's more systematic distinction among different kinds of contingencies. All of these may or may

[75] Several of the essays included in Feinberg, *Doing and Deserving* discuss the problem of luck.

[76] Ibid., 36.

[77] Ibid., 31-2.

[78] Ibid., 37.

[79] This is Feinberg's example, see Feinberg, *Doing and Deserving,* 32.

not be relevant to judgements of moral responsibility including *constitutive luck:* "the kind of person you are[...] [i.e.] inclinations, capacities, and temperament", *circumstantial luck:* "the kind of problems and situations one faces"), and *causal luck:* "luck in how one is determined by antecedent circumstances".[80] Finally, the 'brawl-in-the-pub' case above exemplifies *outcome luck*, i.e. luck as regards the actual consequences of one's actions.[81]

It is an important question whether the immunity-requirement holds equally stringently for all four types of luck and whether these types are really all distinct from one another. In any case, Nagel argues that "all of them present a common problem".[82] Feinberg's own examples also suggest that he believes the ineliminable presence of different kinds of luck to be in equal measure threatening for judgements of moral responsibility.

It may point to deeper differences between Williams's and Nagel's concepts of moral luck that Williams does not share this assumption. He asks: "why do we mind more about it [i.e. luck] in some connection than in others?"[83] Williams is particularly concerned to show that even if it was *metaphysically* possible to meet the immunity-requirement, i.e. even if we could escape 'constitutive luck' and become the unconditioned authors of our actions, the "aim of making morality immune to luck is bound to be disappointed".[84] Several authors have argued since then that the presence of certain kinds of contingent factors may not undermine judgements of responsibility–either because different forms of luck play a different role in our moral judgements or because we cannot make proper sense of certain forms of luck.[85]

The question remains: why do we mind the presence of luck, why do we seek to eliminate it from ascriptions of moral responsibility and why do some find it disturbing that we cannot do so? In the remaining part of this section I would like to reconstruct different arguments that responsibility must be immune to luck. This can help to isolate different explanations for the alleged inconsistency of luck and responsibility. But the main purpose of the following discussion is to argue against those who question the objectivity of moral judgements which do not discount luck. If luck and responsibility

[80] Nagel, 'Moral Luck,' 28ff.

[81] Ibid., 28-32. Zimmerman calls this type of luck *resultant luck,* see his *An Essay on Moral Responsibility,* 38.

[82] Nagel, 'Moral Luck,' 28.

[83] Williams, 'Moral Luck: a Postscript,' 243.

[84] Williams, 'Moral Luck,' 21.

[85] Thus, to cite a few examples from the literature, Hurley has claimed that the idea of 'constitutive luck' is incoherent (see Hurley, *Justice, Luck and Knowledge*), Zimmerman thinks that we must approach resultant luck and situational luck differently (see his *An Essay on Moral Responsibility,* 133-9), Sher holds that constitutive luck about character traits does not undermine blameworthiness (*In Praise of Blame,* esp. 64-6), while Otsuka argues that resultant luck should not be discounted in every case (see his 'Moral Luck Egalitarianism').

are inconsistent it is not because objective judgements of responsibility must discount luck.

Generally speaking, it will be seen that the arguments in defence of the immunity-requirement can be divided into two groups: those which look for moral reasons to explain the inconsistency. These follow the general pattern that if luck is present judgements of moral responsibility are undeserved or unfair. And then there are those arguments which try to demonstrate that the inconsistency is conceptual.

In a number of passages Nagel hints at a possible explanation, namely that it may be *irrational* to take contingent factors on board when making moral assessments.[86] It will remembered that Feinberg makes a similar suggestion when arguing that the rationality of judgements of moral responsibility is limited (certainly inferior to the rationality of judgements of legal responsibility) due to the impossibility of "making precise [their] vaguenesses" and "eliminat[ing] [their] contingencies".[87] Although neither Nagel nor Feinberg elaborate this idea fully (as will be seen shortly the main thrust of Nagel's argument is based on a different consideration), it is clear that this argument seeks to establish the inconsistency of luck and responsibility irrespective of moral considerations on purely conceptual grounds. The inconsistency is said to be rooted in the nature of rational judgements.

It is difficult to see, however, why it would be irrational *not* to discount contingent factors. In fact, we seem to have very good reasons indeed *not* to seek to eliminate luck from our moral judgements. One such reason is that we may simply not be able to do so for epistemic or psychological reasons. Another is that, as several authors have pointed out, we have strong intuitions that luck is not irrelevant after all to moral judgements.[88] "Virtually no one inside or outside the law believes that fault and desert are the sole basis of responsibility", says Honoré for example.[89] Given the strength of these intuitions, there is nothing surprising about the fact that alternative ethical outlooks dispensing with the immunity-requirement have been frequently defended in the past.[90]

The second argument to be considered also rests on the idea of conceptual inconsistency. This is Nagel's main argument and it also appears in Feinberg (although not clearly separated from the third argument, the argument from

[86] See Nagel, 'Moral Luck,' 28, 31.

[87] Feinberg, *Doing and Deserving,* 37.

[88] For instance, Sher appeals to such intuitions as part of his argument that we can be blamed for character traits we cannot help having. He says that we "retain the urge to condemn a miscreant not only for specific acts of cruelty or injustice, but also for the enduring cast of mind that gives rise to these", Sher, *In Praise of Blame,* 60-1. A similar intuition fuels the argument in Robert Adams's 'Involuntary Sins.'

[89] Honoré, *Responsibility and Fault,* 30.

[90] Most importantly, Aristotelian ethics has been described as such an outlook, see Andre, 'Nagel, Williams, and Moral Luck,' 202-7.

objectivity, to be discussed below). It involves equating luck with lack of control. The argument from control is as follows:

1. Whatever is a matter of luck is beyond the agent's control.

   > "There are roughly four ways in which the natural objects of moral assessment are disturbingly subject to luck. All of them present a common problem. They are all opposed by the idea that one cannot be more culpable or estimable for anything than one is for than one is for that fraction of it which is under one's control."[91]

2. Everything that contributes to bringing about an event (including the agent's motives, inclinations, desires and beliefs) turns out under closer scrutiny to be beyond the agent's control.

   > "Everything seems to result from the combined influence of factors, antecedent and posterior to action, that are not within the agent's control."[92]

3. We are only responsible for what is in our control.

   > "So a clear absence of control, produced by involuntary movement, physical force, or ignorance of the circumstances, excuses what is done from moral judgement."[93]

4. Therefore: No one is responsible for anything.

   > "Eventually nothing remains which can be ascribed to the responsible self, and we are left with nothing but a portion of the larger sequence of events."[94]

This is a very powerful argument and one that has been made in similar ways by several authors. One of its apparent strengths is that it seems to do without any appeal to fairness or desert. Nagel's and Feinberg's contribution lies here in showing that one worry about the ineliminability of different kinds of luck reduces to an even more basic worry about the possession of control being an unfulfillable criterion of moral responsibility. The argument provides an explanation of the immunity-requirement by showing that since luck equals control its presence makes moral responsibility metaphysically impossible.

Of course, not everyone accepts this argument. Much of the debate has centered on what kind of control is genuinely required for moral responsibility. Moreover, some have argued that even if we can be shown not to be

[91] Nagel, 'Moral Luck,' 28. For Feinberg's less structured presentation of essentially the same argument, see Feinberg, *Doing and Deserving,* 33-7.

[92] Nagel, 'Moral Luck,' 35.

[93] Ibid., 25.

[94] Ibid., 37.

committed to the condition of control (Premise 3), there is an even more basic argument showing the impossibility of responsibility.[95]

Premise 1 itself is not beyond dispute. There is good reason to think that the *converse* does not hold, i.e. not everything beyond the agent's control is a matter of luck, e.g. events governed by natural laws. But is everything that is a matter of luck beyond the agent's control? Dworkin's concept of *option luck*–"a matter of [...]whether someone gains or loses through accepting an isolated risk he or she should have anticipated and might have declined"[96]–can help us see why this may not always be the case either. There are countless situations in which one opts to accept chancy outcomes. But in such cases the mere fact that the outcome itself was the result of contingent factors does not reduce the agent's moral responsibility for it.

It may be objected here that if the original choice of opting in was itself the result of contingent factors, then the agent's moral responsibility is undermined because the agent has no control over his options. And yet it seems that Dworkin's distinction between option luck and brute luck–"a matter of how risks fall out that are not[...] deliberate gambles"–[97] successfully captures the intuition that luck itself is a complex notion. If that is true, however, then it could be argued that not every of kind of luck impacts in the same way on ascriptions of responsibility. What the distinction shows is that even if everything is, ultimately, a matter of brute luck (because determinism is true and is incompatible with responsibility-entailing control or because determinism is false and indeterminacy is incompatible with responsibility-entailing control), agents make choices and therefore contribute differently to bringing about events than causes of other kinds. In short, even if determinism is true, agents shape the world through their causally efficacious powers differently than impersonal causes.

In any case, I believe that we need not worry too much at this stage about the argument from control for two reasons. *First,* it is a standing assumption throughout this work that there is enough freedom around to make responsibility possible (either because some form of compatibilism or some form of libertarianism is correct). Further, even if that assumption turned out to be unfounded we could still make hypothetical statements as to what the normative implications of ascriptions of moral responsibility *would* be if only we had that kind of freedom. For instance, it seems still quite relevant to inquire why we are prepared to treat option luck differently from brute luck.

[95] This would be the argument from ultimate responsibility which does not require control for moral responsibility. It holds that genuine moral responsibility is nevertheless impossible because agents cannot be the ultimate authors of their actions, G. Strawson, 'The Impossibility of Moral Responsibility.'

[96] Dworkin, *Sovereign Virtue,* 73.

[97] Ibid.

Second, the argument from control renders legal responsibility as well as all normative judgements involving appraisals of agency as vacuous, inapplicable and unjustifiable as judgements of moral responsibility. But even though Feinberg, for instance, does make use of the argument from control, it is manifestly not his view that legal responsibility is on a par with moral responsibility in terms of justifiability. In fact, he treats legal responsibility as a robust concept capable of consistent and valid application. But that is inconsistent with the argument from control. And so it is not surprising that, despite the occasional appearance of the argument from control, it is a different consideration that motivates the skeptical conclusion the Ledger View flirts with. It is that consideration that I would like to discuss next.

The third argument, the argument from objectivity, articulates a worry about the ubiquity of luck that lurks in many discussions of the subject but it is often not made explicit. The crucial idea here is that as long as luck cannot be eliminated from judgements of moral responsibility it will be impossible to guarantee their objectivity and the omnipresence of luck will not allow us to regard judgements of moral responsibility as capable of being true or false.

It will be recalled that according to the Ledger View moral responsibility is to be "read off the facts" leaving no room for discretion or policy. It is because of this requirement that the ineliminability of luck causes difficulties for the Ledger View for the ineliminability of luck raises conceptual barriers to finding a "natural place at which responsibility ends and something else (mere causation?) begins".[98] Only by finding such a natural point for applications of the concept of moral responsibility could we "eliminate its contingencies".[99] Or so it is argued. This concern is of course strongly reminiscent of the skeptical worries raised by the problems of vagueness and causal responsibility for adherents of the Ledger View. Once again, we appear to be left with no alternative but to embrace skepticism about moral responsibility.

I would like to argue, however, that it involves no conceptual inconsistency *per se* to deny the immunity-requirement and simultaneously affirm the justifiability of responsibility ascriptions. It is only that we have to abandon the notion that justifiability can only be guaranteed as long as we succeed in identifying a "natural stopping place" for judgements of responsibility.

In fact, there are a number of positions that deny the immunity-requirement for moral responsibility and yet remain committed to the view that responsibility-ascriptions are judgements. There is, for instance, the radically inflationist view that whether factors (intentions, circumstances, beliefs, etc.) leading to action were subject to luck or not is irrelevant to the

[98] Ripstein, 'Equality, Luck, and Responsibility,' 5.
[99] Feinberg, *Doing and Deserving,* 37.

question of moral responsibility.[100] This view, implausible as it may seem for other reasons, is not incompatible with maintaining that judgements of moral responsibility can be justifiable and objective. One could also argue that the system of "outcome responsibility is the basic type of responsibility in a community[...] a system by which a community allocates responsibility according to outcomes, and we are consequently forced to make bets on those outcomes".[101] This thesis, whatever its merits otherwise, does in no way commit one to the view that responsibility-ascriptions are less justifiable, objective and rational.

Again, the crucial question is why we should move from the claim that there is no "natural" way of circumscribing responsibility to the skeptical conclusion that there is no way of circumscribing responsibility at all. The only reason in sight is the assertion advanced by the Ledger View that moral responsibility is to be "read off the facts" and that blameworthiness/praiseworthiness are "non-moral types of worthiness".

In light of the above discussion it seems that we can run the same argument against the Ledger View for all basic quandaries of moral responsibility: vagueness, causal responsibility and moral luck. In each of these cases, the Ledger View worries that there is no "natural" stopping point for responsibility and that absent such a "natural" point ascriptions of responsibility are vacuous or lack objectivity and are therefore unjustifiable.

But all that goes to show that ascriptions of moral responsibility do not answer to 'brute' facts, not that moral responsibility is vacuous and unjustifiable. The following chapter attempts to sketch such an alternative but cognitivist account of the normativity of judgements of responsibility. There, the appeal to nonmoral worthiness will not play a part and correspondingly what it means to be "true to the facts" in such normative contexts will receive a different interpretation from that of the Ledger View.

5.5 Conclusion

It will be helpful to sum up the criticisms made of the Ledger View in this chapter and also to review some of the insights attributed to this view. What emerged from the above discussion was a conception of "non-moral worthiness" according to which responsibility-ascriptions, if they are to be morally and epistemically defensible, are to track facts 'out there'. By asking whether there is a "natural stopping place" for judgements of responsibility to be found, the Ledger View conjures up a vision of 'responsibility-facts' as part of the natural order of things lending themselves to non-normative description, calling for discovery rather than decision, requiring acute observation as opposed to carefully balanced discretion (which concepts are

[100] This is my reading of Robert Adams's 'Involuntary sins.'

[101] Honoré, *Responsibility and Fault,* 26.

always treated as strict dichotomies). Those sympathetic to the Ledger View are, more often than not, well-aware that such requirements cannot be met. The result is a resigned attitude towards basic quandaries of moral responsibility. These are portrayed as unavoidably undercutting the justifiability and objectivity of responsibility-ascriptions because they exclude the possibility of making judgement-based ascriptions of moral responsibility other than by relying on *ad hoc*, arbitrary and hence morally indefensible rules.

I argued that the principal weakness of the Ledger View was to insist on such implausibly demanding criteria for the justifiability of ascriptions of responsibility. The conclusion with regard to each quandary discussed was either that the difficulty is not restricted to judgements of moral responsibility and not even to normative contexts (as in the case of vagueness), or else, that what the existence of the quandary really shows is that responsibility-ascriptions do not answer to descriptive facts of the kind posited by the Ledger View.

Another shortcoming was found to be the Ledger View's failure to locate responsibility-ascriptions relative to other concerns of morality and practical philosophy. If we were to take a skeptical stance with regard to responsibility-ascriptions, how would that reflect on our judgements about obligations, character traits, practical deliberations, etc.? Would a skeptical conclusion in this specific domain leave the validity of judgements in those other areas intact? We can rightly wonder, I claimed, whether the concept of responsibility being "vacuous" and "inapplicable" would not also impact on one's normative commitments elsewhere.

These negative conclusions can nevertheless prove helpful in exploring the normativity of responsibility-ascriptions. The critical approach of this chapter is by no means intended to give the impression that the Ledger View is of no help whatsoever in confronting these questions. On the contrary, it contains a number of important elements well worth preserving by any cognitivist theory of responsibility.

First, the Ledger View rests on the claim that it is their cognitive component (rather than affective or dispositional attitudes) that is truly distinctive about ascriptions of moral responsibility. In other words, it is because they are based on a judgement that ascriptions of responsibility have normative significance. This claim is of course what qualifies the Ledger View as a cognitivist account of responsibility at the first place (and this is why its skeptical conclusions appear especially threatening). *Second,* the Ledger View is firmly 'retrospectivist' about responsibility. It understands judgements of responsibility as driven primarily by a backward-looking concern about what the agent has done, not what he or other agents are likely to do in the future.

Third, and this is a particularly important finding for the discussion to follow in the next chapter, by insisting on the 'ledger-metaphor', it ties judgements of responsibility closely to the personhood of the agent whom one ascribes responsibility to. Although much of this link between responsibility-ascriptions and personhood remains unexplained within the Ledger View, this 'personalized' approach is based on the valuable insight that judgements of responsibility are motivated by a characteristic interest in *personhood* as a source of what happens in the world and as being categorically different from all other entities with causal powers. And thus *fourth,* the Ledger View recognizes the non-transferability of one's moral responsibility.[102] These are important clues which, even though the Ledger View subscribes to a different understanding of cognitivism, can be valuably exploited in the next chapter which looks to understand the normativity of responsibility.

[102] See for example, Feinberg, *Doing and Deserving,* 136.

Chapter 6

The Value of Responsibility

6.1 Introduction

The task of this last chapter is to make good on the claim that there is a plausible cognitivist, i.e. judgement-based alternative to the theories discussed in the previous chapters. I will argue that the principal plausibility-condition on such a theory is that it must be able to explain what makes responsibility-ascriptions normative and to be able to justify their normative force. In this chapter I will present the Value Thesis in order to show that a judgement-based account is capable of meeting this condition.

Strictly speaking, the Value Thesis is a theory of what makes responsibility normative consisting of the following propositions:

1. Responsibility-ascriptions track the value of being a person capable of recognizing and acting on reasons.

2. The reasons we have for ascribing responsibility are generated directly by our commitment to the value of being a person capable of recognizing and acting on reasons.

3. The reasons we have for ascribing responsibility are independent of our reasons for wanting moral requirements to be met.

4. The judgments which underlie responsibility-ascriptions resemble judgments like x is good, x is cruel, x is worthy, x is beautiful.

5. These judgments respond to a particular kind of value instantiated by a person's ability to recognize and act on reasons.

Accounting for the normativity of ascriptions of responsibility poses a difficulty for any theory of responsibility (I will recapitulate the solutions offered by the theories examined so far in the following section), but it is an especially pressing one for a cognitivist theory. If indeed the primary distinguishing feature[1] of a cognitivist theory is that it holds that an ascription of responsibility involves first and foremost a certain judgement, then the theory must be able to account for that fact that such judgements go beyond a mere description of actions and the agent's relationship to his actions. The Value Thesis is important because, if correct, it can provide such an account, i.e. explain and justify the normative component of judgements which are entailed by any ascription of responsibility.

But is it right to insist that for a theory of responsibility to be plausible it must be capable of accounting for the normativity of responsibility-ascriptions? I think it is, for the following reasons: It is quite clear that ascriptions of responsibility are not made with the intention of merely describing a state of affairs (and not generally regarded as such by those who make them). They matter to us *not* in the way other kinds of descriptions of us matter to us (e.g. 'X is so and so tall') and they do not even matter to us in the way other kinds of evaluative statements of us matter to us (e.g. 'X's nose is beautiful'). It is not so clear, however, how to make sense of and how to justify the special importance we attach to these ascriptions. The talk of the normativity of ascriptions of responsibility is meant to capture why ascriptions matter to us in this special way. They matter to us because they have a special reason-giving force and this is why it is right to call them 'normative'.[2] Thus in Chapter 2, I argued that ascriptions of responsibility

[1] As I explained in Chapter 5, labeling a theory of responsibility 'cognitivist' has more straightforwardly metaethical connotations too. These metaethical implications are also embraced by the judgement-based theory defended in this work, but will only be dealt with in passing in what follows. This is because in order to show conclusively that truth/falsity can be predicated of ascriptions of responsibility, one would need to address broader metaethical considerations which lie beyond the scope of this work. All I was able to do as regards these metaethical concerns was to show in Chapter 5 that nothing in the understanding of responsibility defended in this work excludes that ascriptions of responsibility are indeed capable of being true or false. Later on in this chapter, I will also indicate what it is that, in my view, could make these ascriptions true or false, namely evaluative properties of the agent which he has by virtue of being a person (evaluative properties are picked out by evaluations such as 'x is good', 'bad', 'cruel', 'ugly', etc.). However, I will not be able to show here that these properties are indeed sufficient to make ascriptions of true or false and to guarantee their objectivity. Nor will I be able to address the question how these evaluative properties relate to or supervene on physical properties of the world.

[2] The commonly accepted use of the term 'normative' is this. If something–say an utterance or a state of affairs or a fact–is normative, then it constitutes a reason. This is also how I am using the term above (arguing that responsibility-ascriptions are the source of a special class of reasons for action). Beyond describing this general reason-giving aspect of various things, however, the term 'normative' is also used to refer to the practical motivational power of whatever is called normative, i.e. the psychological potential of say an utterance, state of affairs or a fact to move agents to do or believe or feel one thing or another. Of course, it is a moot question whether recognizing (say) a fact as normative *necessarily* constitutes a reason *for* a rational agent who recognizes it as normative, i.e. whether a normative fact, when recognized as such, necessarily has motivating potential independently of the agent's desires. It is because this is a moot question that one can non-tautologically speak of a reason being normative (=reason-

provide *pro tanto* reasons for action, namely reasons for the imposition of a wide spectrum of normative consequences.[3]

So ascriptions of responsibility matter to us in a special way because they are judgements regarding the appropriateness of *pro tanto* reasons for behaving in certain ways towards agents to whom responsibility is ascribed. If that is true, for a cognitivist theory to be successful it is not sufficient to show that responsibility-ascriptions have an irreducible cognitive component. It must also show that the normative importance of these ascriptions can be explained *in terms of* this cognitive component.

What's more, a cognitivist theory must also say something about the source of the normativity of responsibility-ascriptions. Where does our concern with such ascriptions derive from? Answering this question is indispensable if one is to show that judgements of responsibility are justifiable.

Normative, i.e. reason-giving, considerations can lose their normative force once we have examined the source of their supposed normativity. A command may be normative, it may be a reason for our doing something, but upon realizing who issued the command, we no longer see it as a reason for doing anything. The fact that someone belongs to a certain race is taken by racists to be normative, i.e. as a reason for treating that person in a certain way. Upon realizing the source of our concern with race, the intention to limit social privileges to certain groups, racial classifications will lose (or at least ought to lose) their normative force. They will no longer be (or should be) a reason for anything, but will be reduced at most to descriptive statements. By the same token, it has been argued that the source of our concern with responsibility-ascriptions is a certain conception of Christianity, or ressentiment, or an irradicable vindictive streak in human nature. If any of those claims is true, ascriptions of responsibility will lose their normative force. So beyond the explanatory task of showing why ascriptions of responsibility can be reason-giving, the cognitivist theory also faces the justificatory task of showing that those reasons stand up to scrutiny, that they are consistent with moral or normative concerns of other sorts.

giving). In other words, the expression 'normative reason' is shorthand for saying that something normative, i.e. reason-giving, actually counts for the agent as a reason in favour of a course of action, belief or feeling. With regard to reasons for action, those who deny that a normative fact necessarily counts as a reason for the agent to act accordingly are often referred to as externalists. Externalists deny that something recognized as being normative is necessarily normatively relevant *for the agent's deliberation* as regards what he ought to do. They differ from those who think that something can be normative if and only if it always constitutes at least a *prima facie* reason for action from anyone's deliberative perspective. The latter group includes but is certainly not exhausted by Kantians. A useful classification of theories along these lines can be found in Wallace, 'Normativity and the Will,' 71-81. See also Korsgaard, 'Analysis of Obligation,' 53-4.

[3] As already noted in Chapter 2 (see p. 14), the reasons for action generated by ascriptions of responsibility are only *pro tanto* because they are necessary but not sufficient for the actual imposition of any given normative consequence as most clearly evidenced by the example of punishment.

In the following section, I will discuss how theories of responsibility examined so far account for the normativity of responsibility. This will also be a helpful way of summarizing the findings of previous chapters. I will then go on to present the Value Thesis, my own suggested account of the normativity of responsibility-ascriptions as well as discuss some implications of this thesis and possible objections to it.

6.2 The normativity of responsibility-ascriptions

The main reason for classifying a theory as cognitivist is its commitment to the Priority Thesis, namely the combination of the claims that 'being responsible' is prior to 'holding responsible' and cognitive content is prior to the emotional reactions provoked by the action.[4] I have tried to defend the Priority Thesis in the preceding chapters. In this chapter, I will try to show that the Priority Thesis can be combined with the Value Thesis in order to account for the *normativity* of ascriptions of responsibility, i.e. to explain and justify in terms of judgements the fact that ascriptions of responsibility are not merely descriptions of what someone has done but also provide us with reasons for action.

As already indicated in the introduction to this chapter, in order to gain a better understanding of the normativity of responsibility-ascriptions we need to raise two different types of question. First, we need to know why ascriptions of responsibility can be reason-giving. Second, we need to know whether the reasons we have are good reasons, i.e. whether they are justifiable, by tracing them back to their source. Less abstractly, first, we need to know what sort of concern the reasons generated by ascriptions of responsibility reflect, and second, we need to know whether this concern is itself justifiable. The interest of the first type of question is *explanatory:* 'explain why we hold others and ourselves responsible?'; the interest of the second type of question is *justificatory:* 'is it ever appropriate for us to do so?'. I will now summarize the answers to these two types of question given by theories discussed in the previous three chapters.

The theory examined in the previous chapter, referred to there as the Ledger View, has not given grounds for much optimism as regards the prospects of a cognitivist theory. The Ledger View has been found to be multiply ambiguous as well as mistaken in its principal tenet once the ambiguities have been disentangled. What's more, there is a further problem with the Ledger View, and this is precisely that it seems unable to account for the normativity of ascriptions of responsibility.

What I mean is that we are not explained why saying that there is an entry in someone's ledger of responsibility amounts to more than a mere de-

[4] These claims were defended separately in Chapter 2, see p. 14f. and p. 36. For the formulaic restatement of the Priority Thesis, see Chapter 3, p. 42.

scription. Why should one care whether a certain action occasions an entry into one's ledger? Presumably, we care about being ascribed responsibility for doing wrong because such an ascription entails some kind of condemnation of us as agents. The ledger-metaphor may be useful in bringing out that it is a necessary condition for condemning the agent for an action that that action be inalienably his own (and similarly for praiseworthy actions). However, it remains unclear what the condemnation itself amounts to. Surely, when I object to being condemned in this way I do not simply object to it in the same way as I would dispute (what I believe to be) a false factual proposition about me. I will quarrel with it not only because it is false. But then why do we care about the condemnation (or praise) entailed by an ascription of responsibility and should we care about it all? In general, why does an ascription of responsibility make a difference to how we relate to other people and ourselves?

It only postpones the problem to say that ascriptions of responsibility make a difference because they are necessary to justify the imposition of certain normative consequences on the agent in response to his action (e.g. punishment). It only postpones the problem because then we can ask why it is thought that these ascriptions are required for the justifiability of certain normative consequences and not of others. For example, why do we (tend to) think that only those responsible should be punished?

In contrast to adherents of the Ledger View, few have made a stronger case for the importance of responsibility in our everyday lives than Peter Strawson whose work was discussed in Chapter 4. As regards the *explanatory* question, Strawson's central idea is that we care about ascriptions of responsibility because we "demand a degree of goodwill or regard"[5] from others, especially from those with whom we interact directly, but vicariously even from those with whom we do not. But precisely because we attach such a great importance to the goodwill displayed by others it matters a great deal to us not only what others do to us, but also why they do it, i.e. what attitudes and intentions their actions reflect.[6] Ascriptions of responsibility track these attitudes and intentions establishing whether the action can be traced back to others' goodwill (or the lack of it), or rather the action should not be read as indicative of their attitudes towards us because the action was not voluntarily performed. Hence too the significance of excuses and exemptions. In general, our responses, emotional or otherwise, will be dependent on whether such an ascriptions of responsibility has been made, whether the hurt felt upon someone treading on my hand turns into resentment upon noticing that the other person did so with the intention of harming me.

[5] Strawson, 'Freedom and Resentment,' 7.

[6] See esp. ibid., 5.

As regards the *justificatory* issue, I reconstructed different versions of the Strawsonian answer to the question whether our basic interest in the attitudes displayed by other human beings towards us is justifiable.[7] Common to all of these answers was the claim that for one reason or another–because it is rooted in human nature or because the price of not having it would be prohibitive and so on–this basic interest is inescapable and that fact alone is sufficient to justify it. This response also formed the basis of Strawson's proposed treatment of the challenge of determinism: determinism is irrelevant to the justifiability of ascriptions of responsibility entailed by our reactive attitudes towards others because, although we may have reasons for modifying or suspending attributions of responsibility in particular cases, having recognized their inescapability we need not look further for reasons why we should engage in the practice of responsibility-attributions as a whole.[8]

However, both the explanatory and the justificatory components of the Strawsonian theory suffer from serious shortcomings. As for *explaining* what we do when we ascribe responsibility to one another, I questioned Strawson's claim that reactive attitudes (which are construed by Strawson as responses to the "quality of will" manifest in others' actions) are the key to understanding responsibility. The main reason for my saying so was that the Strawsonian theory seemed to lack the resources to explain what was special about ascriptions of responsibility as opposed to other kinds of reactions to the actions of human beings. The theory could not account for the normative difference between ascriptions of responsibility as opposed to those reactions to others' actions in which goodwill is demanded merely in defense of our own interest and dignity.[9] For example, we had no satisfactory explanation for the difference between resentment and anger–two characteristic responses to the manifest lack of goodwill in others but with quite different normative implications and hence, presumably, quite different conditions of justifiability. I have also argued that the reason why the Strawsonian theory cannot satisfyingly accommodate this difference is because to do so it would have to concede that judgements of responsibility are prior to emotional responses. It is reluctant to make this concession because by conceding this it would also have to give up its claim that our basic interest in other people's attitudes is inseparably tied up with our naturally-ingrained emotional reactions. This link, however, was crucial for the Strawsonian account to explain the special force of ascriptions of responsibility for human beings.

[7] See Section 4.6.

[8] See esp. ibid., 23: "The existence of the general framework of attitudes itself is something we are given with the fact of human society. *As a whole, it neither calls for, nor permits, an external 'rational' justification.*" In Chapter 4, I also noted that arguments from inescapability can be found not only in the Strawsonian theory of responsibility. For example, Tony Honoré seeks to justify our view of other human beings as responsible agents on the grounds that this view is inescapable, see his *Responsibility and Fault,* esp. 30.

[9] Strawson, Freedom and Resentment, 14.

As against this last point, i.e. the attempt to use the alleged inescapability of our basic interest in attitudes voluntarily displayed by human beings to *justify* ascriptions of responsibility, I have argued that although this basic interest may indeed be inescapable, nevertheless this fact does not justify our actual reactions on any given occasion. Even if it is true that in practice we cannot opt out of the practice of attributing responsibility, this practical inescapability will not vindicate the belief that one is rightly held responsible for any given action. Thus should determinism be true (and non-Strawsonian versions of compatibilism wrong), then even if our basic interest in others' goodwill is indeed inescapable, our ascriptions of responsibility will nevertheless be unjustifiable because our beliefs that the attitudes we are concerned with have been *voluntarily* displayed will turn out to be false. In short, what seems irrelevant from a theoretical perspective is not determinism but, quite the contrary, the practical inescapability of certain attitudes.[10]

Consequentialists are quick to identify the reasons we have for ascribing responsibility: ascriptions are made with the purpose of deterring or encouraging certain kinds of actions in the future. This was referred to as the Forward-Looking Thesis in Chapter 3.[11] The answer to what justifies the making of such ascriptions is equally succinct: by deterring or encouraging the kinds of actions for which responsibility is ascribed we maximize expected utility.

This has the notable implication that for consequentialists an ascription of responsibility is justifiable not when the action in question fails to maximize expected utility, but if, and only if, the ascription of responsibility for that action will maximize expected utility. In other words, we have reasons to ascribe responsibility if and only if doing so will maximize expected utility, irrespective of whether the action itself maximized expected utility.[12] It is because of this implication that I said in Chapter 3 that a fully consequent consequentialist must remain skeptical about there being a robust notion of responsibility: an ascription of responsibility on this view will constitute a judgement of the agent only insofar as it is to be judged whether the as-

[10] At the same time, Chapter 4 agreed that inescapability of our basic interest in others' goodwill may be relevant in a different sense: it may be relevant to what we have (practical) reason to do. Even if the inescapability of the basic interest is irrelevant to what it is rational for us to believe, it may be relevant to what it is rational for us to do. For instance, even if determinism is true we may have good rational reasons–"reasons that even outweigh the concern for truth" (Wiggins, 'Towards a Reasonable Libertarianism,' 300)–to behave towards others and ourselves as if it was not. Thus if the truth of some of our theoretical beliefs is irrelevant to what we have reason to do, then we may indeed be faced with a conflict of theoretical and practical rationality. But that understanding is already quite distant from Strawson's original purposes.

[11] See p. 44.

[12] See Sher, *In Praise of Blame,* 123n8. The Influenceability Thesis concerning the relevance of responsibility-undermining excuses and exemptions is a corollary of this principle, see Chapter 3, p. 44.

cription will maximize expected utility or not. But then the problem is that consequentialism also appears to be self-defeating since ascriptions understood in the consequentialist sense are unlikely to exert any influence on the agent himself and are in fact likely to be counter-productive. That is to say, the forward-looking concern of maximizing expected utility seems to be best promoted by ascriptions of responsibility themselves *not* motivated by the forward-looking concern but rather by a robustly retrospective judgement of what the agent has done. But if that is true, ascriptions of responsibility made only with a view to maximizing expected utility lose their point even on the consequentialist account. What's worse, they will therefore become unjustifiable too since, as it was just mentioned, it is part of the consequentialist view that an ascription of responsibility is justified if and only if it will maximize expected utility.

Based on these considerations, the consequentialist position has been extensively criticized both for failing to explain what we do when ascribing responsibility to one another and for not providing an acceptable justification for ascriptions of responsibility either individually or for the practice of responsibility-attributions as a whole. On the one hand, forward-looking considerations just do not seem to be among our reasons for judging that somebody is be responsible for an action (they are at best necessary conditions for holding someone responsible based on that judgement). Nor does it seem to be the case that such reasons *should* be among our reasons when making ascriptions of responsibility. As I said, we lose our grip on the point of making ascriptions of responsibility even when remaining inside the consequentialist framework. When outside that framework, however, it is even clearer that the forward-looking consideration of maximizing expected utility is irrelevant to the justification of responsibility-ascriptions. We attribute responsibility in many cases when the forward-looking is entirely absent, e.g. when the agent cannot be directly confronted with the judgement or is unlikely to be influenced by the judgement.

To sum up, consequentialism misdescribes our interest in the question of responsibility. In fact, quite unlike Strawson's account, it cannot make much sense of why we care about responsibility at all and still less about how deeply we tend to be concerned with our responsibility for what we do. But surely these must be relevant facts about responsibility for any theory. Why should your holding me responsible for something I have done have a special force for me? Why should I quibble over whether you are responsible for something instead of just focusing on whether criticizing you in public would deter you from doing it once again? And why do we have a sense that is it a good thing that people quibble over such things, a good thing, that is, for those people themselves and not only for the promotion of the public good? It is these questions that consequentialists answer wrongly and adherents of the Ledger View lose sight of. And it is these questions which need to be accounted for if the normativity of ascriptions of responsibility is to be better understood. Addressing these questions is what I will attempt to do in the following section.

6.3 Responsibility as value

In this section, I would like to present an alternative way of accounting for the normativity of responsibility-ascriptions in terms of the judgement of the agent that underlies these ascriptions. In Chapter 2, responsibility-ascriptions were defined as normative judgements. Responsibility-ascriptions are normative because they establish reasons, namely *pro tanto* reasons for action. As already argued in Chapter 2, the practical reasons an ascription of responsibility generates are asymmetrical because they are not the same for the addressee(s) of the ascriptions and for others who accept the judgement that underlies the ascription.[13] But it is generally true that a valid ascription of responsibility entails certain things people ought to do.

However, responsibility-ascriptions are frequently taken to be normative in a second, derivative sense too. Apart from their reason-giving force, they are seen as normative also because they are themselves understood to be dependent on norms.

As we have seen, this view is not shared by everyone. Thus according to many, and not only those who subscribe to the Ledger View, moral principles we accept will specify which actions are right and which actions are wrong, but identifying the criteria of the attributability of morally right or wrong actions to agents is not itself dependent on moral or normative considerations.[14] It is claimed that as long as the agent instantiates certain objective properties, e.g. is free from global responsibility-undermining conditions such as severe mental impairment, he will be fit to be held responsible (that is to say, he is capacity-responsible, to use the terminology introduced earlier on[15]) and as long as the action instantiates certain objective properties, e.g. it is voluntarily performed, the agent will be liable to justified praise or blame for it (i.e. be praiseworthy or blameworthy). So facts pertaining to responsibility are straightforwardly natural facts, possibly even physical facts about the agent and the action.

Others argue by contrast that such an 'objectivist' account (of which the Ledger View is just the most fully spelled out example) is wrong because ascribing responsibility to an agent for an action is itself norm-governed. For example, R. Jay Wallace argues that there is no fact of the matter about responsibility "prior to and independent of our moral practices".[16]

[13] See Chapter 2, p. 13.

[14] For the discussion of such 'objectivist' views, see Chapter 2, Section 2.4.

[15] See Chapter 2, Section 2.1.

[16] Wallace, *Responsibility and the Moral Sentiments,* 95.

His claim is not that there are no definite criteria for the attributability of responsibility, but that these criteria are internal to morality and thus dependent on moral norms. The norm Wallace singles out in particular is the moral norm of *fairness.*

It is important to note here that situating the criteria of attributability within the moral domain does not prejudge the outcome of the controversy between compatibilists and incompatibilists. For instance, both camps may and have in fact appealed to considerations of fairness. Thus incompatibilists have argued their case by pointing out that determinism is incompatible with moral responsibility because it is unfair to hold people responsible if determinism is true, while compatibilists (Wallace among them) have objected that the truth of determinism need not have such morally unacceptable implications because it may be fair to ascribe responsibility to agents even if determinism is true.

Whatever we think of the relevance of the appeal to fairness, however, the more general point of interest here is the claim that the justifiability of responsibility-ascriptions is itself governed by norms, norms which are internal to our moral commitments. In other words, the point of departure is the insight that responsibility-ascriptions not only establish reasons but are also made for reasons. The crucial claim is then that these reasons are themselves moral reasons deriving from the moral principles we accept.

This argumentative structure deriving the normativity of responsibility-ascriptions from the acceptance of more general moral principles is fairly widespread. Once again, consequentialism is an important example. As we have seen, consequentialism anchors the reason-giving force of responsibility-ascriptions in the higher-order moral principle requiring us to maximize expected utility. Other examples (including Wallace's own account) will be discussed below.

Tracing back the normativity of responsibility-ascriptions to higher-order moral principles or considerations seems initially plausible. For instance, it is rightly held against objectivist views (such as the Ledger View) that holding someone responsible who suffers from mental impairment–or generally is subject to a valid responsibility-undermining excuse or exemption–is not simply a mistake of fact, but is first and foremost *morally wrong* because cruel or unfair or unjust.

However, I want to question the view that the normativity of responsibility-ascriptions is to be traced back to moral norms. More specifically, I believe there is a case to be made that (i) the normativity of responsibility-ascriptions is to be traced back to the value of responsibility as an aspect of personhood, and (ii) the reasons for making responsibility-ascriptions is generated by our recognition of this value. In other words, the view I would like to defend is that responsibility-attributions are answerable for their appropriateness not to *moral* norms, such as the moral norm of fairness for example, but directly to normatively relevant facts or aspects of the world. It is these normatively relevant facts (not of course particular entities or objects of the natural world, but facts about what is the case) that the appropriateness of responsibility-ascriptions is dependent on. These normatively relevant facts are regarded by human beings as valuable and picked out in our evaluations.

This is not to deny that when we attribute responsibility for a morally wrong (or right) act what we attribute is *moral* responsibility and that our interest in doing is separable from our commitment to morality. But it is to deny that our interest in attributing responsibility is entirely dependent on and derivable from our commitment to moral norms or morality as a whole. I would like to press the view instead that responsibility is something that we also value for its own sake and not only because we recognize the normative importance and priority of moral requirements.

Let me therefore first spell out more formally the argument that links the normativity of responsibility to moral norms, to which I will refer to as the Package Deal Argument, and then present my alternative conception, which I will call the Value Thesis.

6.3.1 The Package Deal Argument and objections

The central claim of the Package Deal Argument is most concisely summed up by Sher (from whom the label for this argument was also taken): "reasons for acting on moral principles and for wanting their requirements not to have gone unmet come as a package deal".[17] This means that an ascription of responsibility is "rendered appropriate by the same considerations–whatever these are–that justify us in accepting the moral principle whose violation, or the disposition to violate which, gives rise to it [i.e. to the ascription of responsibility]".[18] Therefore, "the question 'why blame?' is already implicit in the more familiar 'why be moral?', [so] we may conclude that [...] our reasons for acting morally and for condemning those who do not are indissolubly linked."[19] This is a conclusion that Sher is not alone in endorsing, although others reach it by a somewhat different route (or simply take it for granted).

An influential minority view accepts the Package Deal Argument but concludes that the connection it asserts between the justifiability of moral requirements and ascriptions of moral responsibility, far from securing the justifiability of responsibility-ascriptions, spells trouble instead for the justification of moral requirements. Thus for Bernard Williams the inseparability of morality from a certain notion of blameworthiness jeopardizes the "morality system" as a whole.

[17] Sher, *In Praise of Blame,* 122.

[18] Ibid., 130.

[19] Ibid., 135.

This pessimistic conclusion is reached as follows. Moral responsibility is to track the violation (or meeting) of *moral* requirements according to the Package Deal Argument. But moral requirements are obligations. Obligations obtain categorically. This means that their validity is not contingent on what one needs, desires or wishes to have. Obligations state what there is most reason to do: their normative force is that of unconditioned practical necessity. This entails that obligations for the deliberating agent are both categorically binding and categorically motivating. That is to say, first, obligations for the deliberating agent state what one in any given situation must do irrespective of one's own desires, inclinations and personal commitments. And, second, it also means that one must be motivated to do what one is morally required to do *because* it is morally required (and not because, say, one is inclined that way). For example, to say that the duty to keep a promise is categorical is to say that one ought to keep a promise whether or not one personally benefits from doing so and to say that one ought to keep it *because* it is a duty.

It follows from this that only that part of agency will be relevant to ascriptions of responsibility which is independent from the agent's desires, inclinations and personal commitments. Because ascriptions of moral responsibility track the violation of moral requirements and because moral requirements are meant to be reasons directly valid for the agent (independently from his desires, inclinations and personal commitments), what matters for the justifiability of ascriptions of responsibility is the extent to which the agent was capable of responding to those reasons undetermined by his desires, inclinations and personal commitments. In short, moral responsibility is ascribed only for voluntary actions or voluntarily formed intentions to act. So if the agent was incapable of acting on or was non-culpably unaware of what he was morally required to do, then no moral value whatsoever can be assigned to what he actually does or does not do.

There are two problems with this conception according to the pessimists. First, the notion of categorically motivating reasons is incoherent because I can be only moved to act by reasons which are internalized, i.e. reasons which spring from my own desires and commitments. Second, the notion of a fully voluntary action is also incoherent since, among others, no one voluntarily chooses his character and no action is immune to luck. If that is the case, however, then it appears that no one is ever to be ascribed responsibility for anything he does.[20]

So it seems that morality is left without any adequate way of justifying responses to the violation of its requirements. That raises a difficulty not only for the question whether we can judge agents at all for violations of moral requirements in any given situation. The issue is not merely that of

[20] See esp. Williams, *Ethics and the Limits of Philosophy,* 174-96 and also his 'Moral Luck', 20-1 and 36-9.

the seeming leniency of the 'morality system' in assessing the moral worth of particular actions (which contrasts starkly with the stringency of moral requirements). There is also the more challenging problem concerning the categorical validity of moral requirements. This deeper problem arises because one may doubt the very rationality and justifiability of moral requirements which by definition cannot be met by those to whom it supposedly applies.

It is not at all clear, however, that the success of the Package Deal Argument is predicated either on an incompatibilist understanding of voluntariness or on externalism about reasons, let alone that it is predicated on both.[21] So in what follows, I will not deal with this pessimistic interpretation of the Package Deal Argument. This is also because I believe that there is a more plausible alternative to the Package Deal Argument to explain and justify the normativity of responsibility-ascriptions. I will leave the question open to what extent that alternative is compatible with the obligation-centered understanding of morality discussed above.

But before coming to criticisms of the Package Deal Argument it is worth noting once again that the Package Deal Argument is put forward in order to explain and to justify the normativity of responsibility-ascriptions. Thus what the Package Deal Argument says is, *first,* that the reason-giving force of responsibility-ascriptions is based on our interest in wanting to see moral requirements met. Responsibility-ascriptions generate a special class of *pro tanto* reasons for action for the addressee(s) and addresser(s) of the ascription because they are based on the judgement that the action has violated (or met) a moral requirement. And, *second,* the Package Deal Argument traces back the justification of responsibility-ascriptions to the justification of these moral requirements themselves. Thus the very same reasons we have to accept certain moral requirements as valid will justify us in making ascriptions which single out those who meet those requirements

[21] For a compatibilist and internalist example of a theory which does explicitly subscribe to the Package Deal Argument, see Sher, *In Praise of Blame.* On the other hand, it may be true that for an incompatibilist externalist theory such as Kantian ethics it will be particularly difficult to account for the normative force of ascriptions of responsibility for particular actions. As regards this problem, Korsgaard's insistence on the "generosity of interpretation" of Kantian ethics when it comes to appraising others for their actions is beside the point. It is beside the point because the reasons she cites from Kant in favour of generosity of interpretation apply to the justifiability of 'holding someone responsible' through overt sanctions. These reasons are irrelevant to the question of 'being responsible', *pace* Korsgaard in her 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations,' 205-12. I am not saying of course that this implies anything as regards the Kantian conception of the justifiability and rationality of moral requirements. The point is rather that the price of accepting this conception may be to have to accept that there is no place for blame or praise for particular actions in Kantian ethics. In fact, Korsgaard repeatedly comes close to admitting this, see for example her 'Analysis of Obligation,' esp. 50-1 and 70-1n24, 'Morality as Freedom,' 174 and 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations,' 189 and passim.

and those who do not. In sum, according to the Package Deal Argument morality provides for its own enforcement and simultaneously justifies the enforcement of its requirements.

The Package Deal Argument is seldom set forth in the requisite detail, however. For example, after having dismissed objectivist views as unconvincing, Wallace adopts the already mentioned position that the appropriateness of responsibility-ascriptions is dependent on moral norms.[22] The main consideration presented in support of this view is a negative one concerning the significance of the *absence* of voluntariness, namely that "many of us are tempted, in reflective moments, to think that it would somehow be unfair to treat people as morally responsible if they are deprived of alternate possibilities of action".[23] This consideration, Wallace contends, underlies many incompatibilist arguments too. In other words, the idea is that the appropriateness of responsibility-ascriptions must be dependent on moral norms because it seems that responsibility-undermining excuses and exemptions are themselves adopted for moral reasons. That is to say, we feel that ascribing responsibility to someone who acted under physical constraint or to a child or an insane person is not a mistake of fact or undesirable or impractical, but morally wrong.[24]

But why would the moral principle of fairness require us to suspend our judgement of responsibility if the action was not voluntary? Here the Package Deal Argument kicks in. According to Wallace, to hold someone morally responsible is to view the person as the potential target of a special kind of *moral* appraisal because to hold someone morally responsible is to hold the person to moral obligations (i.e. all the requirements of right/wrong) that one accepts.[25] But moral obligations state binding reasons for action. If, however, the action was not voluntary, then the agent either cannot be said to have acted on a reason at all because he lacked a choice (as in the case of physical constraint) or had no access to the relevant reasons (as in the case of non-culpable inadvertence or mistake). The lack of voluntariness is therefore relevant because (among other admissible excusing conditions such as ignorance) it indicates the "absence of a culpable choice".[26] In sum, the argument is that it is unfair to hold someone responsible whose action was not voluntary *because* judging someone responsible is holding someone to moral requirements we accept but if the action was not voluntary the agent

[22] See Wallace, *Responsibility and the Moral Sentiments,* 91-3.

[23] Ibid., 94. See also the following quote on 98: "[...] we take it as given that we hold people morally responsible–as the result of normal development in cultures that make the reactive emotions available–and we investigate the conditions that make it appropriate to adopt this stance, where the standards of 'appropriateness' appealed to are themselves moral standards."

[24] See ibid., 105.

[25] See esp. ibid., 63-4.

[26] Ibid., 149.

cannot be said to have violated a moral requirement and therefore does not deserve to be held responsible for it.

Already in Chapter 2, I accepted the intuition Wallace points to here arguing that what is wrong with ascribing responsibility for actions not voluntarily performed is that doing so ignores an important feature, namely whether or not people had a genuine opportunity to adjust their behaviour to comply with what was required of them in the given situation. This is unacceptable because agents are persons who deliberate about what they have reason to do, so they are right to feel wronged if they are made to incur responsibility (and the concomitant normative consequences) for actions they did not perform voluntarily.[27]

However, the appeal to that intuition merely explains why it is wrong to ascribe responsibility in the absence of certain conditions such as voluntariness. It does not explain what reasons we have for ascribing responsibility if those conditions *are* met. It may be true that lack of voluntariness excuses because it indicates the "absence of culpable choice" as Wallace says. But this does not explain on the positive side why attributing culpable choice does have normative significance. For all we know, it is even possible at this stage that it is always wrong to ascribe responsibility whether or not the action was voluntary. In other words, the question still is: wherein lies the normative force of the judgement that a person voluntarily flouted or ignored what was required of him? To repeat the questions posed at the end of Chapter 2: Why do we care about being judged to be agents who *voluntarily* acted as reason required? Why do we predicate certain normative consequences on this feature of agency?

The Package Deal Argument serves to answer to these questions. That answer, in short, is that we predicate special reasons for action on the voluntariness of actions because our acceptance of moral requirements or principles commits us to this. By accepting moral requirements as valid, we necessarily accept too that the voluntary infringement of these requirements has special consequences. As we have seen few elucidate the nature of this connection because the Package Deal Argument is often simply taken for granted. But one explanation of the connection runs as follows.[28]

Irrespective of their actual content, all moral principles or requirements display certain formal features including most importantly such features as practicality, universality, omnitemporality as well as (when suitably qualified) overridingness and inescapability.[29] In other words, all moral principles

[27] See Chapter 2, p. 29ff.

[28] What follows is a reconstruction and attempted refutation of Sher's version of the Package Deal Argument, see his *In Praise of Blame,* 123-35.

[29] Note that inescapability is used here in the sense that a valid obligation is not something one can claim not to have to apply to oneself just because one decides not to be a member of the community to whom moral requirements extend. It is this aspect of obligations that Williams refers to when saying that "The moral law is more exigent than

or requirements are practical, that is to say, serve to guide action. Similarly, all moral principles or requirements are universal, that is to say, recognizing them as valid entails recognizing that they apply ”to everyone else who is similarly situated”.[30] And so on for the other features. Although this is not explained we can assume that morality must be said to have these formal features in order to account for that most general characteristic of moral principles and requirements, namely the fact that their normative force is that of unconditioned practical necessity. That is to say, once accepted as valid, they are necessarily accepted as categorically binding too.

Now, the crucial claim on this presentation of the Package Deal Argument is that these formal features ”rule out the possibility of fully accepting a moral principle without wanting those who have ignored or flouted its requirements not to have done so and those who are disposed to ignore or flout its requirements not to be so disposed”.[31] The basic idea[32] is that because all moral requirements are prescriptive accepting moral requirements amounts to adopting a ”favourable attitude toward whichever action it prescribes”.[33] From this it follows, it is claimed, that accepting moral requirements as valid necessarily entails our wanting everyone to act in accordance with those moral requirements in the past as well as in the future. Therefore, as long as the moral requirements are justified, we are *pro tanto* justified too in imposing normative consequences on agents who do not act in accordance with those moral requirements.

My principal objection to the Package Deal Argument is that ignores our interest in the question why agents act the way they do and therefore ignores our interest in agents as persons. Our responses to agency, because agents are persons, reflect a special interest not so much in what agents do, but rather in what their reasons were for acting the way they did. Or at least ascriptions of responsibility reflect such an interest and this why they presuppose voluntariness of the action. To put the point more formally, the argument above cannot accommodate the fact that the justifiability of our (pro tanto) reasons to act in response to agents depends on the voluntariness of their action.

It may be true, even analytically true as Sher says,[34] that given the formal features of moral principles or requirements we are committed to wishing for the non-occurrence of their infringement in the world. It may

the law of an actual liberal republic, because it allows no emigration”, Williams, *’Ethics and the Limits of Philosophy,’* 178. I used the term ’inescapability’ somewhat differently in Chapter 4. The term there referred to the claim that concern with agency and responsibility is rooted in human nature.

[30] Ibid., 125.

[31] Ibid., 124.

[32] For the introductory discussion of this idea, see Chapter 2, p. 39.

[33] Sher, *In Praise of Blame,* 124.

[34] Ibid., 133.

be true, that is, that we do not want to be and want others to be people who Φ-ed, if Φ-ing is wrong, and we do not want this *because Φ-ing is wrong.* However, this does not explain why we do not want to be and want others to be people who *voluntarily* Φ-ed, if Φ-ing is wrong (and why we want to be and want others to be people who *voluntarily* Φ-ed, if Φ-ing is right). The normative significance of voluntariness and *a fortiori* the normative significance of responsibility-ascriptions does not follow from the above mentioned formal features of moral principles or requirements.

By adopting the Package Deal Argument we would come too close for comfort to the consequentialist justification of responsibility-ascriptions. Although, unlike in consequentialism, it is not claimed here that responsibility-ascriptions are motivated by an exclusively forward-looking concern about what happens in the world, it is claimed that responsibility-ascriptions are motivated by a concern about what happens in the world past and future, namely a minimum of infringements of moral principles and their requirements. We take a special interest in agency on this account because agency is a source of moral rightness or wrongness in the world and we wish moral rightness maximized and wrongness minimized.

The problem is not restricted to Sher's presentation of the Package Deal Argument. I believe that it spreads to all attempts which seek to derive the normativity of responsibility-ascriptions from the practical necessity of moral requirements.

Note that there are two basic ways to account for the practical necessity of moral requirements. One is to say that moral requirements are binding because one will be ascribed responsibility for violating them. That is, "the primary force of saying that I am obliged to do something is that I will be judged, punished, blamed or will blame myself, if I do not".[35] The other is to derive the bindingness of moral principles or requirements from the formal features of morality (which formal features need to be validated by some independent and prior procedure).[36] As we have seen, Sher's presentation of the Package Deal Argument was based on this understanding of the bindingness of moral principles and their requirements.

For the first method to avoid circularity, it needs to supply additional arguments to explain how our interest in being judged, punished and blamed and in judging, punishing and blaming others–in short, our interest in the attributability of responsibility–is based inescapably and universally in human

[35] See Korsgaard, 'Analysis of Obligation,' 50. Korsgaard traces back this account of obligation to the moral sentimentalism of Hume and Hutcheson. But it can be found in Williams's work too. See for example his *'Ethics and the Limits of Philosophy,'* 180: "But obligations have a moral stringency, which means that breaking them attracts blame." And also ibid., 177: "Blame is the characteristic reaction of the morality system."

[36] Kant's analysis of the conceptions of good will and rational action can be seen as such a procedure to secure such formal features for morality and to derive the bindingness of moral principles and their requirements from these formal features. See Korsgaard, 'An Introduction to the Ethical, Political, and Religious Thought of Kant,' 12-18.

nature. Such additional arguments are necessary because once one says that an action is binding if an agent incurs responsibility for failing to perform it, one cannot simultaneously say that being responsible just is to be held to the violation of binding requirements. In other words, if (the normative force of) bindingness is defined through responsibility, then (the normative force of) responsibility cannot be defined through bindingness. The insistence of Strawsonian theories of moral responsibility on the inescapability and universal rootedness in human nature of the "basic demand" for good will can be seen as precisely such an attempt to establish the naturalistic and therefore independent basis of the normative significance of responsibility-ascriptions. Wallace too, for example, says in this vein that: "There is a primitive desire to be thought well of by others, which is refined through socialization into a second-order interest in justified regard: the desire not to act in ways that could incur the well-ground resentment or indignation of others."[37] However, attempts to justify responsibility-ascriptions on the basis of their alleged inescapability are unpersuasive for reasons discussed in Chapter 4.

As for the second method, I believe that what has been said in connection with Sher's presentation of the Package Deal Argument applies to all attempts to derive the normativity of responsibility-ascriptions from the formal features of moral requirements. Even if it is accepted that, owing to these formal features of morality, moral principles or requirements are categorically binding, it does not follow from this that we have categorically binding (pro tanto) reasons for imposing normative consequences on those who violate (or meet) these requirements. In other words, from the fact that a certain course of action is obligatory from the perspective of the deliberating agent (by virtue of the formal features of the moral requirement that calls for that course of action), it does not follow that we have reasons, let alone categorical reasons, from the third-person perspective to impose normative consequences on the basis of the agent's responsibility for not choosing that course of action. More precisely, our justified interest in moral requirements themselves does not explain why we should be concerned with the question whether the agent chose *voluntarily* to violate (or meet) those requirements or not and why we should have special reasons for action if indeed the violation (or fulfillment) was *voluntary.*

6.3.2 The Value Thesis

The Value Thesis presents an alternative to the Package Deal Argument, which if the above claims are correct, is sorely needed. The Value Thesis contends that ascriptions of responsibility track a value, the value of being a person capable of recognizing and acting on reasons. Unlike objectivist

[37] Wallace, *Responsibility and the Moral Sentiments,* 70.

views, the Value Thesis does not think of 'being responsible' as a natural fact about persons. On the other hand, it also differs from those conceptions which are based on the Package Deal Argument. This is because it claims that the normative significance of responsibility lies not merely in our commitment to not wanting the requirements of morality go unheeded, a commitment which, according to the Package Deal Argument, is entailed by our commitment to moral requirements themselves. According to the Value Thesis, the normativity of responsibility-ascriptions is to be traced back to what we value about people. If that is true, responsibility-ascriptions generate reasons for us not only insofar as they trace the violation of moral requirements. Rather, their normativity has to do with the value of responsibility as an aspect of personhood which is the value of (i) being a person with the general capacity of recognizing and acting on reasons, and (ii) being a person who has on a given occasion recognized and acted on reasons.

Therefore, the Value Thesis differs from the Package Deal Argument at a number of critical junctures.

First, it denies that the reasons we have for ascribing responsibility are the same reasons we have for accepting certain moral requirements or principles as valid (whatever these requirements or principles may be). Instead, the Value Thesis proposes to see the reasons we have for ascribing responsibility as generated *directly* by our commitment to the value of being a person capable of recognizing and acting on reasons. Hence it denies that the very same considerations that justify us in accepting a given moral principle justify us in ascribing responsibility. Rather, the Value Thesis traces back the normativity of responsibility-ascriptions to the value we attribute to being a person capable of recognizing and acting on reasons.

Second, the Value Thesis holds that the normativity of responsibility-ascriptions is not dependent on the normativity of moral requirements or principles. That means that judgements concerning the responsibility of others and ourselves have action-guiding significance for us not because we accept certain moral requirements or principles and we do not want to have those requirements or principles unmet. We attach value to the feature of personhood which enables persons to recognize and act on reasons independently from what the content of these reasons may be. So our answer to the question 'why ascribe responsibility?' will not be entailed by our answer to the question 'why be moral?'.

On this conception, moral responsibility is only one, though certainly important, application of the general concept of responsibility. In principle, this even leaves open the logical possibility of generally recognizing ascriptions of responsibility as normative, while denying that ascriptions of moral responsibility are normative–that is, if one wanted to deny the normativity of moral reasons altogether. Or less strongly, the possibility that ascriptions of responsibility are on a par with ascriptions of moral responsibility in terms of their reason-giving force, i.e. that ascribing responsibility for the violation or meeting of moral requirements does not produce stronger reasons for imposing certain normative consequences than ascriptions of responsibility for actions which violate or meet other kinds of norms. Whether one wants to embrace these options depends on one's assessment of the claim that moral reasons are overriding and that they constitute a separate class of reasons at the first place.[38] I cannot take a stand on either of these issues here. My point is only that even if we do regard moral considerations as authoritative in a special sense, the normativity of responsibility-ascriptions should not be seen as deriving from this authority. It is compatible with this point to say, however, that ascriptions of responsibility for voluntarily flouting or ignoring moral reasons have special normative implications because of the strength of these reasons.

And it follows, *third,* that according to the Value Thesis the judgements which underlie responsibility-ascriptions are the closest to evaluative judgements of the kind: 'x is good', 'cruel', 'just', 'worthy' and even 'amusing', 'disgusting', 'beautiful'. In other words, when making a judgement that ascribes responsibility to an agent for an action, we respond to a particular kind of value instantiated by a person's actualization of his ability to recognize and act on reasons.

The significance of the third point emerges when we ask what the appropriateness of responsibility-ascriptions depends on. We have seen that those who embrace the Package Deal Argument criticized objectivist views for the implausible assumption of the latter that there is a realm of brute facts 'out there' on which one's responsibility for an action depends. The alternative offered was that the standards appealed to determine the appropriateness of ascriptions of responsibility are themselves moral standards.[39] The Value Thesis broaches the possibility of a third option, namely that ascriptions of responsibility are answerable neither to brute natural facts nor to internal norms of morality but rather to a substantive value, that of being a person who is able to recognize and act on reasons. To put it differently, the Value Thesis has affinities with but also differs from both the objectivist position such as the Ledger View and the position of theories endorsing the Package Deal Argument. It differs from the latter and is closer to the former position in that it holds responsibility-ascriptions directly answerable for their appropriateness to an aspect of the world (rather than norms internal to morality), but it differs from the former position and closer to the latter in that it sees this aspect not as constituted by brute natural facts but rather by substantively valuable properties of personhood.

Does the Value Thesis require us therefore to revise the claim that unless an action is voluntary the agent cannot be ascribed responsibility for

[38] On these issues, see Raz, 'On the Moral Point of View.' Raz denies that morality covers a domain of reasons which hang together as a class. But cf. Wallace, 'The Rightness of Acts and the Goodness of Lives,' for criticisms of this view.

[39] See above and Wallace, *Responsibility and the Moral Sentiments,* 98.

it? Specifically, I endorsed in Chapter 2 the claim that voluntariness is necessary for blameworthiness (and praiseworthiness).[40] But some rely on the Package Deal Argument to support this claim. Thus, as we have seen, it is argued that it is some moral principle, for example the principle of fairness, which requires us to recognize voluntariness as a necessary condition for responsibility to be imputable to the agent. Unless the action was performed voluntarily, it would not be *fair* to hold the agent responsible because unless the agent could do otherwise holding him responsible is undeserved and hence unfair.[41] It is therefore a higher-order moral principle, here the principle of fairness, that requires us to abstain from judgement of responsibility unless the action was voluntary in the required sense.

I do not accept the Package Deal Argument. So do I have to rescind the claim the voluntariness is necessary for responsibility? No, because I believe one can agree with the claim that it can be wrong, because cruel or unjust or unfair, to hold someone responsible for an action not voluntarily performed without trying to account for the wrongness of doing so by appealing to the normativity of moral requirements. If it is indeed unfair to hold someone responsible for an action not voluntarily performed, it is not because the lack of voluntariness indicates an inability (temporal or global) to meet *moral* requirements, but because of something more general.

In addition, it is also important to note the following: The argument in favour of the Value Thesis has so far been *negative* appealing to the Package Deal Argument's failure to satisfyingly anchor the normativity of responsibility-ascriptions in the normativity of moral requirements. But the Value Thesis also draws positive support from the consideration that we ascribe responsibility not only for the violation of moral requirements, but for other things people voluntarily do as well. I believe that we ascribe responsibility for *voluntary* violations (or fulfillments) of many important norms which loom large in our lives and which cluster around central values of aesthetics, decorum, rationality as well as around central values of other domains. And thus conversely too, in many cases it is considered to be unfair to judge someone for an involuntary action which violated a requirement *not* of morality but that of aesthetics or decorum or rational self-interest or instrumental reason.

In order to show that this is the case, however, I would have to analyze the nature of each of these norms separately and show what kind of practical reasons they generate and why. I cannot undertake such an analysis here as it would involve a detailed examination of each of these normative domains. Therefore, I will limit myself to the consideration of a single example where, arguably, the norm in question is *not* moral and yet we are held responsible for meeting it or failing it voluntarily. If the example is successful, then

[40] See p. 29ff.

[41] See above and Wallace, *Responsibility and the Moral Sentiments,* 106-7.

we have found at least one case (and quite an important one I believe) in which ascriptions of responsibility have special normative force for us despite the fact that responsibility is not in this case ascribed for the violation or meeting of a moral requirement. That is positive evidence in favour of the Value Thesis because it demonstrates that the reason-giving force of responsibility-ascriptions is not dependent on our commitment to moral principles and their requirements.

The example is about that special kind of choice which involves "adopt-[ing] a particular plan of life."[42] That kind of choice need not have moral aspects or implications. In any case, the crucial point for me here is that that kind of choice need not have moral aspects or implications and yet we can be 'called to task' for it. If by no one else, then by ourselves. In other words, it is not meaningless to say that the life-plan we have adopted did not cause harm to anyone or cause a setback to anyone's interests and yet to say that it was the *wrong* choice. There need not be moral cost involved (or the moral cost may not be so high as to affect the moral justifiability of the choice itself) in order for the choice to be objectionable.

For instance, it has been argued that the choice of life-plan can be subject to criticism by others (and by myself at a later stage of my life) on the grounds that "it violates the principles of choice" or that it did not carefully evaluate the consequences or that it was made in ignorance of some of the relevant facts.[43] Alternatively, one may reject the idea that "the model of rational deliberation" can be applied to one's choice of a life-plan.[44] But even if one rejects that idea, the choice of life-plan can be assessed in terms of important but *non-moral* criteria or norms. Thus we can applaud Gauguin's choice of life-plan in Williams's celebrated example for having led to the production of great works of art. And we can also applaud Gauguin's choice for being consequent, for being reflective in the sense of based on self-knowledge, for deciding to pursue a meaningful life and so on. Conversely too, as countless examples from life and literature demonstrate, we can criticize life-choices (including our own) which embrace mediocrity, easy success, etc. at the expense of pursuing meaning, commitment and engagement.[45]

The point is then that such life-choices are instances of voluntary agency for which we are accountable. Whether they are positive or negative, we are ascribed responsibility for them (often by ourselves) and these ascriptions of responsibility too have normative force in the sense that they generate reasons justifying actions towards or by the person who took a particular life-choice. If indeed we accept Josiah Royce's thought that "a person may be

[42] Rawls, *A Theory of Justice,* 415.

[43] See ibid., 408-9.

[44] This is Williams's criticism of Rawls in his 'Moral Luck', see esp. 33ff.

[45] On the good of engagement and how engagement constitutes an irreducible and indispensable way of relating to value for human beings, see Raz, *Value, Respect, and Attachment,* esp. 162-3.

regarded as a human life lived according to a plan",[46] then one's choice of life plan will be central to many of our interactions with and attitudes towards other persons and ourselves. It follows that ascriptions of responsibility for such life-choices, though potentially morally neutral, have great practical significance.

If that is true, we have at least one important example which demonstrates that our normative interest in the voluntariness of actions is not derived from or dependent on our commitment to moral obligations. In addition, I believe that examples from other normative domains could be developed to underscore the same point, even though for reasons of space and complexity I cannot do so here.

It follows from these considerations that the acceptance of the Value Thesis does not force us to rescind the claim that voluntariness is a necessary condition of responsibility. On the Value Thesis the ability to act voluntarily is not valued because it renders us capable of acting in conformity with moral requirements. Conversely too, our willingness to excuse involuntary actions has to do with our interest in people's capacity to recognize and act on reasons (because this is an aspect of personhood we value), rather than merely our interest in seeing moral obligations discharged. In sum, being capable of voluntary action is being responsive to and capable of acting on reasons, properties of personhood we value for their own sake.

Unconvinced by the positive and negative arguments in favour of the Value Thesis made in the foregoing, one could continue to object to the the Value Thesis as follows: Is it not true that *moral* responsibility is attributed for the violation or fulfilling of *moral* requirements? And doesn't the normative force of such an ascription derives from the fact that what the agent violated (or fulfilled) were specifically *moral* requirements?

Suppose, for instance, that Φ-ing is wrong. Surely, in that case, ascribing moral responsibility to X for Φ-ing also says something about X himself, i.e. that he was the kind of person who voluntarily violated a moral requirement. So one could argue that ascriptions of moral responsibility are linked to a special class of reasons for action because the ascription reflects on the moral worth of the agent as a person. Is this not sense in which responsibility-ascriptions are taken to be evaluative of personhood? Don't we want to say, contrary to the Value Thesis, that the agent's blameworthiness implies that by doing wrong he has voluntarily failed the norms or principles of morality we accept and this is where the action-guiding significance of responsibility-ascriptions is really derived from?

The worry about the Value Thesis expressed in this objection can also be put as follows. Responsibility-ascriptions typically give rise to certain kinds of reasons for action, i.e. reasons to impose normative consequences in response to the blameworthy/praiseworthy action such as punishment.

[46] Quoted in Rawls, *A Theory of Justice,* 408.

But the normative consequences, for the imposition of which ascriptions of responsibility furnish us with reasons, appear to be inevitably consequential upon the violation or fulfilling of moral requirements. For example, the judgement that an agent has violated a valid moral requirement furnishes us with a *pro tanto* reason to punish him, to morally censure him or to express our resentment and so on. It seems to be a necessary feature of these reactions that our explanation and justification of them will "invoke a moral concept and its associated principles".[47] Doesn't the Value Thesis commit one to denying what therefore appears to be a necessary connection?

I do not think so. What the Value Thesis denies is that our reasons for ascribing responsibility are explained and justified by our reasons for accepting certain moral requirements as valid, i.e. that it is our acceptance of certain moral requirements that commits us to ascribing responsibility because by accepting the moral requirement we are necessarily committed to wanting to see those moral requirements fulfilled. I believe it is compatible with this position to say that when a moral requirement is violated it is indeed the violation of the moral requirement that we ascribe responsibility for.

As I said above, this would be a special, though certainly important application of the general concept of responsibility. It is an important application because an important class of reasons is constituted by moral requirements. Valuing responsibility as an essential aspect of personhood means that we also value people's ability to recognize and act on all kinds of reasons, including the specifically *moral* reasons which moral requirements give rise to. Quite likely too, this particular ability has special value for us, therefore we acknowledge the existence of reasons for imposing special normative consequences when a person's action does not appear to respond to or respect this value, i.e. when the action violates moral requirements we accept. So it may well be true that by accepting a moral requirement as valid we are necessarily committed to wanting to see that moral requirement fulfilled *and this is why we ascribe responsibility for it, if it is not fulfilled.* Where the Value Thesis differs is in insisting that ascribing responsibility is explained and justified by our concern to see moral requirements fulfilled, that we value responsibility because we value moral requirements.

6.3.3 Defending the Value Thesis

What remains to be seen is how well the Value Thesis fares in meeting the tasks of explaining and justifying the normativity of responsibility-ascriptions relative to its competitors. Is the Value Thesis capable of explaining what we care about when we ascribe responsibility and can it justify our concern? Thus, more formally, *first,* we have to see whether the Value

[47] Rawls, *A Theory of Justice,* 481.

Thesis is capable of explaining why we hold others and ourselves responsible. In other words, we need to ask whether we view ascriptions of responsibility as reason-giving *because* we value the ability of persons to recognize and act on reasons. Do the normative consequences predicated upon valid ascriptions of responsibility indeed reflect our commitment to the value of responsibility as the Value Thesis conceives of this commitment? *Second,* we need to know whether our commitment to the value of responsibility as an aspect of personhood justifies our ascriptions of responsibility. In other words, the question is whether it is ever appropriate to ascribe responsibility on the basis of our commitment to the value of responsibility. It may be true, as the Value Thesis claims, that the source of the normativity of responsibility-ascriptions is indeed our commitment to the value of responsibility as an aspect of personhood. But is this commitment capable of justifying the reasons generated by responsibility-ascriptions? Should we care about responsibility as a value and are the reasons generated by the commitment to this value cogent and binding?

I believe that the Value Thesis does well in explaining the good to be had from engaging in the practice of responsibility-attributions.[48] Responsibility, moral or otherwise, is frequently portrayed as burdensome, as something placed on the agent's shoulders and the practice of holding one another responsible for our actions as downright cruel and vindictive.[49] But this portrayal already loses much of its plausibility once the different senses of responsibility, as distinguished by Hart, are recalled.[50] It will be seen then that these portrayals disregard the great importance we attach to what Hart calls capacity-responsibility, i.e. being mature persons who deliberate and reach decisions autonomously and free from external influence such as brainwashing and who are also able to carry out these decisions once made.[51] This capacity is something that has great value and the status it guarantees

[48] Consequentialism often talks explicitly of the "good of blaming", see Chapter 3. But consequentialism has been criticized for misrepresenting this good by describing the value of responsibility-ascriptions to consist exclusively in their contribution to maximizing the expected utility of future actions through deterrence and encouragement. Strawsonians also appeal to the value of reactive attitudes and feelings predicated on ascriptions of responsibility, see Chapter 4, p. 109ff. But here the commitment to this value is claimed to be naturally ingrained and therefore inescapable. The Value Thesis does not regard the commitment to the value of responsibility as either naturally ingrained or inescapable.

[49] For a particular vigorous presentation of this view, see Wertheimer, 'Constraining Condemning,' and Baier, 'Moralism and Cruelty: Reflections on Hume and Kant.' It is unclear whether the same criticism would be made by these authors of the conception of responsibility which, as the Value Thesis, does not trace back the normativity of responsibility-ascriptions to the normativity of moral requirements.

[50] Hart, 'Postscript: Responsibility and Retribution.' See Chapter 2, Section 2.1 for a detailed discussion of the various senses of responsibility.

[51] Kant's essay 'Was ist die Aufklärung?' is one of the classic defences of the value of this capacity.

is something people can often be seen to be reluctant to part with even if they could gain materially by doing so.[52]

As we have seen, some authors highlight the normative importance of capacity-responsibility by pointing out that it would be wrong (not just mistaken) to impose normative consequences such as punishment on an agent who lacks this capacity (due to hypnosis, serious mental impairment, etc.). But the normative importance of this capacity is also reflected in the fact that having this capacity is something we value for its own sake. Typically, it is considered to be a serious offence to deny that an adult person is capacity-responsible and such claim must always rest on substantial evidence before it is allowed to influence our judgement of that person. Also, the absolute prohibition on torture, for example, is partly justified by the fact that it forcefully deprives a person of his status as a responsible person. However, this latter aspect of the normativity of capacity-responsibility, i.e. the intrinsic good of being (and being regarded as) a responsible person, is insufficiently articulated by conceptions which, accepting the Package Deal Argument, seek to derive the normative importance of responsibility-ascriptions from our acceptance of moral principles and their requirements.

It can be objected here that while the general capacity enabling one to take responsibility for one's action may indeed be something valuable and to that extent the Value Thesis may be persuasive, it is hard to explain how ascriptions of responsibility in particular cases can be said to represent a commitment to the value of responsibility. Specifically, the problem is to account for the reason-giving force of ascriptions of responsibility in particular cases in terms of the commitment to the value of responsibility. Why would the value we generally attach to being a person capable of recognizing and acting on reasons furnish us with *pro tanto* reasons to impose certain normative consequences on an agent who is held responsible for a particular action? According to the Package Deal Argument we have such reasons because the action violated (or met) some moral requirement we accept. That would also explain why ascriptions of responsibility and the imposition of concomitant normative consequences are justified even though they in many cases do not appear to serve in any way the future good of either the person judged or of those judging him.

What can the Value Thesis say about this objection? I believe that the Value Thesis can respond to this objection by characterizing in more detail the value of responsibility as an aspect of personhood. Thus the value we attach to being a person capable of recognizing and acting on reasons is

[52] This insight forms the basis of John Gardner's account of the differing implications of standard defenses in court. As he persuasively explains, while both the defenses of provocation and diminished responsibility may lead to substituting a charge of manslaughter for a murder conviction, the price of the latter is to lose the status of a responsible, self-respecting person who had good reasons to do what he did. See Gardner, 'The Gist of Excuses,' esp. 590-2.

what Raz has described as an enabling value. Enabling values are "values whose good is in making possible or facilitating the instantiation of other values."[53] The value we attach to responsibility as an aspect of personhood is such an enabling value because by committing ourselves to this value we can relate to one another in certain distinct ways, namely as persons capable of recognizing and acting on reasons. So instead of merely serving to promote or bring about or enforce moral requirements, responding to the value of responsibility enables us to give a distinct shape to human relationships. That explains well why we value the capacity-responsibility for by valuing this capacity we can relate to other people (and ourselves) as reasons-responsive beings whose interactions are based on the exchange of reasons.

But I believe it also explains the normativity of responsibility-ascriptions in particular cases. As I argued in Chapter 2 already, acting puts the agent under "justificatory pressure".[54] This is one immediate normative consequence of action-attribution (with or without the agent's responsibility). That means that an ascription of responsibility is also a request addressed at the agent to give the reasons for, i.e. to justify, his action. A judgement that the agent is blameworthy (because he is responsible for an all-things-considered unjustifiable action[55]) is a judgement to the effect that ultimately such reasons cannot be found. What, if any, other normative consequence, such as punishment, will be imposed on the basis of that judgement will, as we have seen, depend on a range of further considerations.[56] Among others, it will depend on the strength and character of the reasons the agent had not to act the way he did. If we accept the view that a valid moral requirement (unless defeated by a stronger one and unless the agent is non-culpably unaware of the existence of that reason) always constitutes an overriding reason for the agent, then the agent's not having acted in compliance with that requirement will constitute a reason to impose specific normative consequences such as moral censure or punishment.

6.4 Responsibility as an aspect of personhood

I have repeatedly used the expression that we are committed to the value of responsibility, more specifically to the value of being a person capable of recognizing and acting on reasons. But in this chapter and the foregoing ones I have criticized the view according to which our commitment to the prac-

[53] Raz, *The Practice of Value,* 34.

[54] See Chapter 2, p. 34.

[55] See definition on p. 14 in Chapter 2.

[56] For instance, punishment involves harsh treatment and requires authorization of the person or body inflicting the punishment, so additional reasons must be supplied to justify the claim that the agent is liable to be punishment rather than (say) overt censure. See Chapter 2, p. 20.

tice of responsibility-attributions is inescapable. It may even be true that human nature predisposes us to be responsive to the value of responsibility, i.e. relating us to other human beings as capable of recognizing and acting on reasons, but that is no proof of the correctness or justifiability of our judgements that anyone is ever responsible for anything. Or so I claimed. But if the putatively natural roots of this commitment are irrelevant to its justification, then we have to ask, as mentioned above, whether or not our commitment to this value is itself justifiable.

So the task in this final section is to explain why being responsible, i.e. being a person capable of recognizing and acting on reasons, is an aspect of personhood to be valued. But, as criticisms of Strawsonian theories of responsibility show, the sought-after explanation cannot be that our commitment to the value of responsibility as an aspect of personhood is naturally ingrained therefore inescapable (because that explanation cannot show how inescapability can justify ascriptions of responsibility). This suggests a different way of explaining why being a responsible agent is an aspect of personhood to be valued. This explanation would be that being a responsible agent is a valuable aspect of personhood because being a responsible agent is *constitutive* of what it is to be a person.

According to this explanation, ascriptions of responsibility are expressions of the status of persons and having the status of being a responsible agent is itself valuable. Note, however, that a full explanation along these lines would require working out a theory of personhood. Only after having such a theory in hand could one explain fully the sense in which being responsible is constitutive of personhood. This is because an array of further questions would have to be answered first: How does being a responsible agent as an aspect of personhood relate to other properties–such as inviolability or being ends-in-themselves or being unconditionally valuable or having a rational will–on which the status of personhood is often taken to rest? In particular, does the claim that being responsible is constitutive of personhood entail that being a responsible agent is *necessary* for having the status of personhood (or, if having the status of a person is something that comes in degrees, that one is a person to the extent one is a responsible agent)? This would imply that one could not be inviolable or an end-in-oneself or unconditionally valuable or have a rational will unless one was a responsible agent. I do not think that having any of these properties requires that one is a responsible agent. For instance, the status of inviolability does not seem to be affected by the fact that one is not capacity-responsible (due to serious mental impairment, say). On the other hand, there is a good case to be made that being a responsible agent is *sufficient* for having the status of personhood. That is to say, one can be a person without being a responsible agent, but one cannot be a responsible agent without being a person.

Even if this is correct, however, to fully explain why being a responsible agent is a valuable aspect of personhood is something that, as I said, cannot be done without a theory of personhood. Working out such a theory is beyond the scope of this work. What I can do here is to indicate two, perhaps mutually compatible, views both of which go some way towards explaining why responsibility is a valuable aspect of personhood.

One important view about the value of responsibility is that assuming and attributing responsibility is crucial to our identity as persons. This view has been defended, among others, by Tony Honoré who says that "if actions and outcomes were not ascribed to us on the basis of our bodily movements and their mental accompaniments, we could have no continuing history or character".[57] It is because we ascribe actions and decisions to persons as authors of these actions and decisions that persons have identity and character. But it is by virtue of their identity and character that persons have a special status. So by defining personal identity, ascriptions of responsibility are constitutive of the status of persons: "as the counterpart of this status we are responsible for our actions and their consequences".[58] In short, "to be responsible is part of what it means to be a person".[59]

It is important to distinguish the notion of personal identity at issue here from the metaphysical problem of the numerical identity of persons over time. *That* problem raises the question of what are the necessary and sufficient conditions for a person existing at time t^1 to be the same person as the person existing at t^2. But it is also important to distinguish the notion of personal identity invoked by Honoré from the psychological or everyday usage where talk of one's identity refers to one's most central values, concerns or pursuits. In fact, one could say that the notion of personal identity under scrutiny here is located in between these two other concepts. It supervenes on the metaphysical identity of persons (to the extent there is such a thing) and it makes possible the having of personal identity in the psychological or everyday sense just mentioned.

Thus the personal identity which ascriptions of responsibility are constitutive of, I submit, is the identity of a person who lives his life according to a plan (which plan can of course be subject to change due to both objective and subjective reasons and which plan can take on ever more concrete forms as one 'goes along'). Royce's suggestion already referred to above[60] is helpful here. According to Royce, an individual establishes his identity, "says who he is", by describing his purposes, pursuits and his future plans for his life. I contend then, following Honoré, that ascriptions of responsibility are constitutive of personhood insofar as personhood is to choose a life-plan and to execute it, i.e. to live the plan that one has chosen.

[57] Honoré, *Responsibility and Fault,* 29.

[58] Ibid., 29.

[59] Ibid., 30.

[60] See p. 169 above. Royce is discussed by Rawls in *A Theory of Justice*, 408.

Human beings could not be persons in this sense without being responsible agents for a number of reasons. *First,* because we could not sensibly talk of choosing a plan and living in accordance with it if we did not regard ourselves as capable of voluntary actions based on our deliberations. We adopt such plans in order to answer the question "what to do with our life".[61] That presupposes that it is up to us what we do with our lives. Up to us, in the sense that our voluntarily chosen actions will shape the course of our lives to the extent permitted by external circumstances. And up to us also in the sense that we 'can be called to task' for both the choice itself (I have already talked about this[62]) and our success in carrying out the plan we have chosen. The leeway permitted by external circumstances may often be severely limited. But it would not make sense to make such plans if there was no leeway at all. And because there is leeway, responsibility can be constitutive of personhood.[63]

Second, if it is true that happiness depends to a great extent "on the successful execution (more or less) of a rational plan of life drawn up under (more or less) favourable conditions",[64] or even if it is at least true that happiness depends on the successful execution of some plan, whether rational or not, then it is also true that happiness will also depend on the extent to which what one has turned out to be has been a function of one's voluntary choices. This is not to say that luck cannot make one happy, not at all. But it is to say that the knowledge that what one has achieved is at least in part due to one's own voluntarily chosen actions can increase happiness and fulfillment. And this is precisely why one regards responsibility as a valuable aspect of one's personhood.

In sum, on the interpretation of Honoré's suggestion I propose here, responsibility is constitutive of personhood whereby personhood is understood as living life according to a plan. If that is true, then the Value Thesis is compatible with the general claim that the value of responsibility derives from attributions of responsibility being constitutive of personal identity. Linking responsibility to the identity of persons would be one way of anchoring the value of responsibility as the value of being a person capable of recognizing and acting on reasons. But the Value Thesis diverges from Honoré's position in a number of important ways.

First, Honoré believes that responsibility for outcomes is the fundamental type of responsibility as opposed to responsibility which presupposes fault (he refers to the latter type as *moral* responsibility). That is to say, Honoré does not share the view that voluntariness is a necessary condition

[61] Ibid., 413.

[62] See p. 169 above.

[63] Although, as I said above (p. 175), I think we would be persons even if there was no leeway. That is to say, responsibility is constitutive of but not necessary for being a person.

[64] Ibid., 409.

of the basic, personal identity constituting type of responsibility. This latter contention is questionable. In my opinion, the value of responsibility can be argued to derive from the constitutive importance of responsibility-attributions for personal identity only if voluntariness is preserved as a necessary condition of identity-constituting responsibility (while it may still be true that outcome responsibility is the fundamental notion in some well-circumscribed areas such as the law of torts). This is because, as Stephen R. Perry has observed, "outcomes contribute to our identity precisely because we are responsible for them".[65] Accordingly, the most plausible thing to say is that the crucial link between outcomes and personal identity is precisely our *voluntary* contribution to bringing about those outcomes. What we voluntarily do is what is definitive of our identity as persons and not just any outcome of our actions.

Second, Honoré holds that the basic type of responsibility, which is constitutive of the identity and therefore of the status of persons, is inescapable.[66] There is no need to repeat here the criticisms previously made of arguments from inescapability. Whether or not attributions of responsibility (be it outcome or any other kind of responsibility) are inescapable, the inescapability of the practice will not justify an ascription in any given instance.

Perhaps recognizing the force of this objection, Honoré also adds, *third,* that although to be responsible is inescapably part of what it means to be a person, it is also in our interest to ascribe (outcome) responsibility to one another and ourselves because we are more likely to benefit in the long run from a fair system of allocating (outcome) responsibility. On this conception, when we choose a course of action we bet on outcomes (the payoffs being social credits and discredits). Honoré concedes that on this conception our responsibility and therefore the running total of our credits and discredits will inevitably depend on luck, but if the rules of allocating responsibility are drawn up in a fair way (i.e. they are "impartial, reciprocal, and over a period, beneficial"[67]), then we stand to gain from such a social arrangement which is therefore overall justifiable.

But for one thing note that this "social" justification of outcome responsibility is totally different from the conception that takes attributions of responsibility to be constitutive of personhood. It is unlikely that these two understandings–the "personhood understanding" and the "social understanding"–are compatible with one another even if Honoré is sympathetic to both.[68] The value of attributing responsibility (whether outcome

[65] Perry, 'Honoré on Responsibility for Outcomes,' 71.

[66] See for example Honoré, *Responsibility and Fault,* 30.

[67] Ibid., 26.

[68] As persuasively argued in Perry, 'Honoré on Responsibility for Outcomes' where the two understandings are carefully prized apart and their incompatibility demonstrated, see esp. 63, 66.

responsibility or some other form of responsibility) can derive from it being constitutive of personal identity or from it being a mutually beneficial social arrangement. If it is the latter, it must be in some sense chosen or endorsed by persons and therefore cannot also be taken to be constitutive of personhood.

Moreover, it is unlikely that a system of allocating outcome responsibility would be chosen because it is unlikely that it can be designed in a way that will benefit everyone concerned even in the long run.[69] In fact, what we encounter here is once again the view already criticized in connection with rule-consequentialist theories of responsibility. According to this view the good of responsibility is supposed to derive from the fact that the practice of ascribing responsibility makes possible an optimal social arrangement, optimal because it maximizes expected utility (this is the consequentialist position) or because it serves everyone person's interests best (Honoré's position). The Value Thesis opposes this view not simply because the purported explanation of why we attribute responsibility to one another seems to be empirically false. It also objects that the good of responsibility derives not from grounding a specific kind of social arrangement, but rather by making various kinds of social arrangements possible at the first place. This is because ascriptions of responsibility enable people to relate to one another in distinct ways, namely as persons capable of recognizing and acting on reasons. On the Value Thesis this, if anything, is what justifies the normativity of responsibility-ascriptions rather than the practice of responsibility-ascriptions constituting a specific social arrangement which is alleged to be most likely to serve people's interests or promote their welfare.

Another feasible way of tracing the value of responsibility back to personhood is offered by contractualism, at least in the version of it defended by Thomas Scanlon. On this account, the value of responsibility has to do with the role played by ascriptions of responsibility in enabling people to participate in inter-*personal* relationships. In other words, instead of focusing on how responsibility-ascriptions ground personal identity, this view emphasizes the constitutive role of responsibility-ascriptions for people engaged in a "system of co-deliberation" based on a "fully reciprocal recognition of one another"–on a "kind of idealized reciprocity of respect".[70]

[69] See again Ibid., 67: "We all know people, or know of people, who apparently posses whatever minimum capacity is required to get by in the world and be properly regarded as a person, who nonetheless seem to be (and to be destined from the outset to be) life's perennial losers."

[70] Insofar the criticism made in the preceding paragraph of the "social" justification of the practice of responsibility echoes the contractualist approach. I will try to explain below how the Value Thesis diverges from this approach. The expression "system of co-deliberation" occurs among others in Scanlon, 'The Significance of Choice,' 166, while the terms "fully reciprocal recognition of one another" and "idealized reciprocity of respect" are taken from Wallace, 'Scanlon's Contractualism,' 282.

The principal tenet of contractualism is that an act is morally wrong "if its performance under the circumstances would be disallowed by any set of principles for the regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement".[71] Therefore, as long as people are motivated to act morally they will seek to regulate their behaviour to comply with "standards that others could not reasonably reject insofar as they, too, were looking for a common set of principles".[72] It follows that people, as long as "suitably motivated", will be primarily concerned with the question whether their actions could be justified to others (who are moved by the same concern).

So the notion of interpersonal justification, or "justifiability to others", takes central stage on this account: what we owe to others is that our actions remain justifiable to others in terms of mutually acceptable principles. Consequently, ascriptions of responsibility serve two main purposes. First, they are understood as judgements assessing whether the agent acted in accordance with moral requirements generated by such principles. Second, they call upon the agent to reconsider the attitudes expressed in his action and modify or withdraw the attitudes if this was not the case.

Thus on this version of contractualism the normativity of responsibility-ascriptions–the particularly strong moral criticism they express–comes down to their special "significance for our relations with a person".[73] Violations of moral requirements are tantamount to disregarding other people because they involve acting in a way that would not pass the test of interpersonal justifiability. Doing wrong is to flout the legitimate demand others have on us to regulate our actions in compliance with mutually acceptable principles. The value of responsibility lies in the contribution of ascriptions of responsibility to maintaining this "system of co-deliberation" in which moral criticism and moral argument "consist in the exchange of requests and justifications".[74] Ascriptions of responsibility establish who the appropriate targets of moral criticism and moral argument are, namely people capable of having and acting on "judgement-sensitive attitudes", i.e. people who can make judgements about what there is reason to do and who can act on such judgements as well. Also, as noted above, they track whether actions are in accord with mutually accepted standards, and when they are not they call upon the agent himself to recognize this fact.[75] In sum, on the contractualist account, most people care about responsibility because "most people

[71] Scanlon, *What We Owe to Each Other,* 153.

[72] Scanlon, 'The Significance of Choice,' 166.

[73] Scanlon, 'Reasons, Responsibility, and Reliance: Replies to Wallace, Dworkin, and Deigh,' 511.

[74] Scanlon, 'The Significance of Choice,' 171.

[75] In many cases, of course, the demand for recognition will not or cannot or should not be made overt. But it is always implicit in the judgement of responsibility, cf. esp. Scanlon, 'Reasons, Responsibility, and Reliance,' 512.

care about the justifiability of their actions to others"[76]–the normativity of responsibility is derived from the requirement of interpersonal justifiability.

The value of responsibility lies therefore in its being a constitutive aspect of personhood which, on this version of contractualism at least, is itself defined in essentially interpersonal terms. A Scanlonian contractualist would certainly agree that being a person is partly to possess the capacity for critically reflective, rational self-governance, which capacity can be exercised even without recognizing that one owes anything to others. But beyond this capacity of 'intrapersonal reasons-responsiveness', being a person is also constituted by a kind of interpersonal responsiveness to the demands made upon one by the existence of other persons. As long as we are interpersonally responsive we will "exercise our capacity for self-governance in ways that others could reasonably be expected to authorize".[77]

For contractualists, we are tempted to say, being a person is a relational property. Existing as a person is impossible without the existence of other persons. Ascriptions of responsibility play a crucial role in making clear, in addition, that for being a person it is not only the existence of other persons that matters but also our active recognition of others as persons. We value responsibility as an aspect of personhood, that is, because ascriptions of responsibility spell out, first, to whom we owe justification for our actions, namely persons who are capable of judging what there is reason to do, and second, because ascriptions of responsibility tie us into the system of co-deliberation by calling on us to reflect on our reasons for action ensuring that they remain justifiable to others in terms of mutually acceptable principles.

I hoped to make clear through the above presentation of the contractualist grounding of the value of responsibility in its understanding of personhood the affinities between contractualism and the Value Thesis. At the same time, the approach outlined in the Value Thesis differs from the contractualist account for the following reasons.

First and most important is the objection that the contractualist account also relies on the Package Deal Argument which has been criticized above. In other words, if we accept the Value Thesis, we may still agree with the part of the contractualist conception which says that we value responsibility, i.e. being a person capable of recognizing and acting on reasons, because it enables us to relate to other people in certain distinct ways. But I believe we can reject the other central contractualist claim about responsibility, namely that we value responsibility only because it enables us to relate to other people as co-deliberators of *moral* principles. We are interested in the reasons-responsive (judgement-sensitive) attitudes of other people not only because we are interested in people's responsiveness to moral reasons as specified by mutually acceptable principles.

[76] Scanlon, 'The Significance of Choice,' 170.
[77] Scanlon, 'The Significance of Choice,' 174.

In fact, it seems to me that this objection is made even more plausible by Scanlon's readiness to recognize the great normative importance of moral and non-moral values beyond the morality of right and wrong (it is to the latter domain that Scanlon applies the contractualist method). These values include friendship, parenthood, honour, integrity, achievement, and sex.[78] Are we not also interested in how people respond to the reason-giving potential of these values? Don't we value responsibility also because it tracks the ability of persons to respond to and act on the reasons these values may generate?

Second, whatever we think of the notion that justifiability to others is central to thinking about what is morally right and wrong, we have reasons to question the corollary that appears to follow from this notion, namely the contention that we care about responsibility only because we care about justifying our actions to *others.* It seems plausible to say that we care about responsibility also because we care about justifying our actions to ourselves. In fact, this aspect of the value of responsibility is focused on by Honoré as we have seen above. That is to say, the normativity of ascriptions of responsibility derives not only from their constitutive role in establishing our interpersonal status as co-deliberators with other members of the moral community, but also from being constitutive of one's own identity as a person.

If that is true, then the good of responsibility for the individual is constituted not only by the significance of responsibility as aspect of personhood for "our continued relation"[79] to other people who may display various attitudes towards us. But ascriptions of responsibility are equally significant for our continued relation to ourselves. The point here is not that the contractualist account outlined above cannot account for the importance self-reflexive ascriptions of responsibility when one holds oneself responsible for an action.[80] What the objection says is that ascriptions of responsibility made by others as well as oneself determine one's view of oneself beyond the issue whether one perceives one's reasons for action to be justifiable to others.

At the end of the present section, I have turned to two different accounts of how the value of responsibility may be grounded in some conception of personhood. Although certain aspects of both theories appear to be problematic, the criticisms I have made of them do not touch on the central contention common to both, namely that responsibility is valued because it is constitutive of personhood. The most contentious assertion entailed by the Value Thesis (which also distinguishes it from the contractualist account of why responsibility is valued as an aspect of personhood) is that

[78] See esp. Scanlon, *What We Owe to Each Other,* 171-7.

[79] Scanlon, 'Reasons, Responsibility, and Reliance', 512.

[80] For examples of the contractualist analysis of such self-reflexive cases, see Scanlon, 'The Significance of Choice,' 167, 171-2.

regarding ascriptions of responsibility as normative does not presuppose our prior commitment to some set of moral principles and their requirements. The justification of responsibility-ascriptions does not depend on and is not derived from some set of moral principles we accept. The value we attach to responsibility is in this sense pre-moral and instead of being anchored in morality it is anchored in what we take a person to be.

Bibliography

Adams, Robert Merrihew. 'Involuntary Sins.' *Philosophical Review* 94 (1985): 3-31.

Andre, Judith. 'Nagel, Williams, and Moral Luck.' *Analysis* 43 (1983): 202-7.

Arneson, Richard J. 'The Smart Theory of Moral Responsibility and Desert.' In *Desert and Justice,* edited by Serena Olsaretti. Oxford: Oxford University Press, 2003.

Ayer, A. J. 'Free-Will and Rationality.' In van Straaten, *Philosophical Subjects,* 1-13.

Baier, Annette. 'Moralism and Cruelty: Reflections on Hume and Kant.' In *Moral Prejudices: Essays on Ethics.* Cambridge, MA: Harvard University Press, 1995.

Bennett, Jonathan. 'Accountability.' In van Straaten, *Philosophical Subjects,* 14-47.

Dennett, Daniel. *Elbow Room.* Cambridge, MA: MIT Press, 1984.

Dodds, E. R. 'On Misunderstanding the Oedipus Rex,' *Greece and Rome* 13 (1966): 37-49.

Dworkin, Ronald. 'No Right Answer?' In *Law, Morality and Society,* edited by P. M. S. Hacker and Joseph Raz. Oxford: Oxford University Press, 1977.

———. 'Objectivity and Truth: You'd Better Believe It,' *Philosophy and Public Affairs* 25 (1996): 87-139.

———. *Sovereign Virtue.* Cambridge, MA: Harvard University Press, 2000.

Farkas, Katalin. 'Szkepticizmus és Filozófiai Gondolkodás,' *Világosság* 41 (2000): 59-78.

Feinberg, Joel. *Doing and Deserving.* Princeton: Princeton University Press, 1970.

———. *Rights, Justice, and the Bounds of Liberty: Essays in Social Philosophy.* Princeton: Princeton University Press, 1980.

———. 'The Moral and Legal Responsibility of the Bad Samaritan,' In *Freedom and Fulfillment.* Princeton: Princeton University Press, 1992.

Gardner, John. 'The Gist of Excuses,' *Buffalo Criminal Law Review* 2 (1998): 575-98.

———. 'In Defence of Defences.' In *Floris Juris et legum. Festskrift till Nils Jareborg,* edited Peter Asp, Carl Erik Herlitz, Lena Holmqvist. Uppsala: Iustus Frlag, 2002.

——— and Timothy Macklem, 'Reasons.' In Jurisprudence and Philosophy of Law, edited by Jules Coleman and Scott Shapiro. Oxford: Oxford University Press, 2002.

Gibbard, Allan. *Wise Choices, Apt Feelings: A Theory of Normative Judgement.* Oxford: Clarendon Press, 1992.

Glover, Jonathan. *Responsibility.* London: Routledge and Kegan Paul, 1970.

Hart, H. L. A. 'The Ascription of Responsibility and Rights,' In *Proceedings of the Aristotelian Society* 49 (1949), 171-94.

———. *The Concept of Law.* Oxford: Clarendon Press, 1961.

———. 'Positivism and the Separation of Law and Morals.' In *Essays in Jurisprudence and Philosophy* (1958). Oxford: Oxford University Press, 1983.

———. 'Postscript: Responsibility and Retribution.' In *Punishment and Responsibility.* Oxford: Clarendon Press, 1968.

Honderich, Ted. *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes.* Oxford: Oxford University Press, 1988.

Honoré, Tony. *Responsibility and Fault.* Oxford: Hart Publishing, 1999.

Hume, David. *A Treatise of Human Nature* (1739-40), edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Oxford University Press, 1978.

Hurley, Susan. *Justice, Luck and Knowledge.* Cambridge, MA: Harvard University Press, 1993.

Johnson, Conrad D. 'The Authority of the Moral Agent.' In *Consequentialism and its Critics,* edited by Samuel Scheffler. Oxford: Oxford University Press, 1988.

Kant, Immanuel. 'Was ist die Aufklärung? (1783)' In Weischedel, ed. *Kants Werke,* Bd. 9, 53-61.

———. *Die Metaphysik der Sitten (1798).* In Weischedel, ed. *Kants Werke,* Bd. 7, 303-634.

Keefe, Rosanna and Peter Smith. *Vagueness.* Cambridge, MA: MIT Press, 1997.

Korsgaard, Christine M. *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press, 1996.

———. 'An Introduction to the Ethical, Political, and Religious Thought of Kant.' In Korsgaard, *Creating the Kingdom of Ends,* 3-42.

———. 'Kant's Analysis of Obligation: The Argument of *Groundwork I.*' In Korsgaard, *Creating the Kingdom of Ends,* 43-76.

———. 'Morality as Freedom.' In Korsgaard, *Creating the Kingdom of Ends,* 159-88.

———. 'Creating the Kingdom of Ends: Reciprocity and Responsibility in Personal Relations.' In Korsgaard, *Creating the Kingdom of Ends,* 188-221.

———. 'The Reasons We Can Share: An Attack on the Distinction between Agent-Relative and Agent-Neutral Values.' In Korsgaard, *Creating the Kingdom of Ends,* 275-310.

Latus, Andrew. 'Moral Luck.' In *The Internet Encyclopedia of Philosophy.* http://www.iep.utm.edu/m/moralluc.htm.

Moody-Adams, Michele. 'On the Old Saw that Character is Destiny.' In *Identity, Character, and Morality: Essays in Moral Psychology.* Edited by Owen Flanagan and Amelie Oksenberg Rorty. Cambridge, MA: MIT Press, 1990.

Moore, Michael S. 'Causation and Responsibility,' *Social Philosophy and Policy* 16 (1999): 1-51.

Murphy, Jeffrie. 'Moral Death: A Kantian Essay on Psychopathy,' *Ethics* 82 (1972): 284-98.

Nagel, Thomas. 'Moral Luck.' In *Mortal Questions.* Cambridge: Cambridge University Press, 1979.

Nowell-Smith, P. H. 'Freewill and Moral Responsibility,' *Mind* 57 (1948): 45-61.

———. *Ethics.* Baltimore: Penguin, 1954.

Otsuka, Michael. 'Moral Luck Egalitarianism.' Paper given at the Bled Philosophical Conference (Slovenia), June 2006.

Perry, Stephen R. 'Honoré on Responsibility for Outcomes.' In *Relating to Responsibility: Essays in Honour of Tony Honoré on his 80th Birthday,* edited by John Gardner and Peter Cane. Oxford: Hart Publishing, 2001.

Pettit, Philip. *A Theory of Freedom: From the Psychology to the Politics of Agency.* Oxford: Oxford University Press, 2001.

Rawls, John. 'Two Concepts of Rules (1955).' In *John Rawls: Collected Papers,* edited by Samuel Freeman. Cambridge, MA: Harvard University Press, 1999.

———. *A Theory of Justice.* Oxford: Oxford University Press, 1971.

Raz, Joseph. *The Authority of Law.* Oxford: Clarendon Press, 1979.

———. 'Legal Positivism and the Sources of Law.' In Raz, *The Authority of Law,* 37-52.

———. 'Legal Reasons, Sources, and Gaps.' In Raz, *The Authority of Law,* 53-77.

———. 'The Functions of Law.' In Raz, *The Authority of Law,* 163-232.

———. *Practical Reason and Norms.* Oxford: Oxford University Press, 1990.

———. 'On the Moral Point of View.' In *Engaging Reason: On the Theory of Value and Action.* Oxford: Oxford University Press, 1999.

———. *Value, Respect, and Attachment.* Cambridge: Cambridge University Press, 2001.

———. *The Practice of Value.* Oxford: Clarendon Press, 2003.

Ripstein, Arthur. 'Equality, Luck, and Responsibility,' *Philosophy and Public Affairs* 23 (1994): 3-23.

Russell, Paul. 'Strawson's Way of Naturalizing Responsibility,' *Ethics* 102 (1992): 287-302.

Sartorio, Carolina. 'Omissions and Causalism.' Paper given at the Massachusetts Institute of Technology's Meeting of the Minds, January 27, 2007. http://web.mit.edu/philos/www/mm/sartorio.pdf.

———. 'How To Be Responsible For Something Without Causing It.' In *The Oxford Handbook of Causation,* edited by Helen Beebee, Christopher Hitchcock and Peter Menzies. Oxford: Oxford University Press, forthcoming. http://philosophy.wisc.edu/sartorio/ce.doc.

Scanlon, Thomas. 'The Significance of Choice.' In *The Tanner Lectures on Human Values,* Salt Lake City: University of Utah Press, 1988.

———. *What We Owe to Each Other.* Cambridge, MA: Harvard University Press, 1998.

———. *The Difficulty of Tolerance.* Cambridge: Cambridge University Press, 2003.

———. 'Reasons, Responsibility, and Reliance: Replies to Wallace, Dworkin, and Deigh," *Ethics* 112 (2002): 507-28.

Schlick, Moritz. *Problems of Ethics,* Chapter 7. New York: Prentice Hall, 1939. Reprinted as 'When is a Man Responsible?' In *Free Will and Determinism,* edited by Bernard Berofsky. New York: Harper and Row, 1966. Citations are to this edition.

Sher, George. *In Praise of Blame.* Oxford: Oxford University Press, 2006.

Smart, J. J. C. 'Freewill, Praise, and Blame,' *Mind* 70 (1961): 291-306.

Stern, Lawrence. 'Freedom, Blame, and Moral Community,' *Journal of Philosophy* 71 (1974): 72-84.

van Straaten, Zak, ed. *Philosophical Subjects: Essays Presented to P.F. Strawson.* Oxford: Oxford University Press, 1980.

Strawson, Galen. *Freedom and Belief.* Oxford: Clarendon Press, 1986.

———. 'The Impossibility of Moral Responsibility,' *Philosophical Studies* 75 (1994): 5-24.

Strawson, P. F. 'Freedom and Resentment.' In *Freedom and Resentment.* London: Methuen, 1974.

———. 'Reply to Ayer and Bennett.' In van Straaten, *Philosophical Subjects,* 260-96.

———. *Skepticism and Naturalism. Some Varieties.* London: Methuen, 1985.

Swinburne, Richard. 'The Christian Scheme of Salvation.' In *Philosophy and the Christian Faith,* edited by Thomas V. Morris. Notre Dame: University of Notre Dame Press, 1988.

Thomson, Judith. 'Imposing Risks.' In *Rights, Restitution, and Risk: Essays in Moral Theory.* Cambridge, MA: Harvard University Press, 1986.

———. 'Causation: Omissions,' *Philosophy and Phenomenological Research* 66 (2003): 81-103.

von Wright, Georg Henrik. *Norm and Action.* New York: The Humanities Press, 1963.

Wallace, Jay R. *Responsibility and the Moral Sentiments,* Cambridge, MA: Harvard University Press, 1994.

———. *Normativity and the Will.* Oxford: Clarendon Press, 2006.

———. 'Normativity and the Will.' In Wallace, *Normativity and the Will,* 71-81.

———. 'Moral Responsibility and the Practical Point of View.' In Wallace, *Normativity and the Will,* 144-64.

———. 'Scanlon's Contractualism.' In Wallace, *Normativity and the Will,* 263-99.

———. 'The Rightness of Acts and the Goodness of Lives.' In Wallace, *Normativity and the Will,* 300-21.

Watson, Gary. 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.' In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology,* edited by Ferdinand Schoeman. Cambridge: Cambridge University Press, 1987.

Weischedel, Wilhelm, ed. *Immanuel Kants Werke in zehn Bänden.* Darmstadt: Wissenschaftliche Buchgesellschaft, 1983.

Wertheimer, Roger. 'Constraining Condemning,' *Ethics* 108 (1998): 489-501.

Wiggins, David. Towards a Reasonable Libertarianism. In *Needs, Values, Truth,* 3rd revised ed. Oxford: Oxford University Press, 2002.

———. *Ethics,* London: Penguin, 2006.

Williams, Bernard. 'Moral Luck.' In *Moral Luck.* Cambridge: Cambridge University Press, 1981.

———. *Ethics and the Limits of Philosophy.* Cambridge: Harvard University Press, 1985.

———. *Shame and Necessity.* Berkeley: University of California Press, 1993.

———. 'Moral Luck: A Postscript.' In *Making Sense of Humanity and Other Philosophical Papers.* Cambridge: Cambridge University Press, 1995.

Williams, Michael: *Unnatural Doubts.* Princeton: Princeton University Press, 1996.

Wolf, Susan. 'The Importance of Free Will,' *Mind* 60 (1981): 386-405.

Zimmerman, Michael J. *An Essay on Moral Responsibility.* Totowa: Rowman and Littlefield, 1988.

———. 'Rights, Compensation, and Culpability,' *Law and Philosophy,* 13 (1994): 419-50.

———. *The Concept of Moral Obligation*, Cambridge: Cambridge University Press, 1996.