Responsibility as Attributability:
Control, Blame, Fairness

By

Anna Réz

Submitted to
Central European University
Department of Philosophy

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Philosophy

Supervisor:
Prof. Ferenc Huoranszki

Budapest, Hungary

2013

To the best of my knowledge this thesis contains no materials previously written or published by any other person, except where it was so indicated. This thesis contains no material accepted for the award of any other degrees in any other institution.

Budapest, May 2013 Anna Réz

Abstract

Attributionism is a fairly new type of theory of moral responsibility. In his influential book *What We Owe to Each Other* Thomas Scanlon distinguished two senses of responsibility, *substantive responsibility* and *responsibility as attributability* and provided a nuanced description and analysis of both concepts. Elaborating on the latter notion in a series of articles Angela Smith developed a unified account of attributonism, the “rational relations view”. According to Neil Levy’s formulation, “on the attributionist account, I am responsible for my attitudes, and my acts and omissions insofar as they express my attitudes, in all cases in which my attributes express my identity as a practical agent. Attitudes are thus expressive of who I am if they belong to the class of *judgment-sensitive attitudes*” (Levy 2005).

One of the main advantages of attributionist accounts is that they are able to explain and justify some puzzling cases of responsibility: responsibility for attitudes and responsibility for involuntary omissions. These cases are troubling because they reveal an inconsistency in our ethical thinking: on the one hand, we seem to be committed to an important moral principle, the Control Principle, which states that it is unfair to hold people responsible for things beyond their control. But, on the other hand, with our ordinary judgments of responsibility we frequently assess people on the basis of such things over which they do not exercise control.

In my thesis I wish to accomplish a dual aim. First, I give a comprehensive and thorough analysis of attributionist theories. I explore how they differ from apparently similar accounts, the strengths and weaknesses of their solutions to traditional problems of moral responsibility. I raise several objections and investigate whether attributionist accounts have the resources to answer them. Although I do not attempt to defend attributionist theories from every criticism, hopefully I can demonstrate that attributionism has several appeals which make it a genuine rival of more traditional accounts of moral responsibility.

Second, exploring attributionist accounts serves more general purposes. The analysis, as I indicated, will lead us to the discussion of the Control Principle. I explore the problem emerging from the principle and give an abstract mapping of the possible solutions for it. One of these strategies lead us to the discussion of R. J. Wallace's much debated normative interpretation, which claims that one is morally responsible for something if and only it is fair to hold her responsible—facts about responsibility are defined by normative considerations regulating the fairness of responsibility-attribution. The normative interpretation, put forward as a general schema, has far-reaching methodological consequences. Most importantly, as I will argue, any theory of responsibility has to define three variables: the scope of responsibility-attribution, the nature of the relevant responsibility-attributing practices and the substantive moral considerations about fairness which should be applied. Thus, in the second part of the thesis I will explore these topics as they arise for attributionist theories. Also, the normative interpretation raises fundamental and often neglected questions about the methodology of building up a theory of free will and responsibility and the division of labor between theories of responsibility and substantive normative ethical theories. At the end of the discussion I focus on these questions and try to clarify some important methodological issues which often remain implicit in the relevant literature.

Table of Contents

Introduction

The Free Will Debate – A Very Short Introduction

The problem of free will and moral responsibility has a long and respectful history in the philosophical literature. Traditionally the issue was a dominantly metaphysical one, asking how the world should be in order to make free and responsible agency possible. In particular, standard standpoints were shaped by two questions, concerning the truth of causal determinism (a concept subject to many debates and in need of further clarification) and its compatibility with freedom of the will. Commonly we distinguish three positions characterized by how they respond to these two questions. While *libertarians* give a negative answer to both questions, trying to make room for undetermined, thus free and responsible actions, *hard determinists* and their contemporary followers, *hard incompatibilists* claim that moral responsibility, given the metaphysical structure of our world, is impossible[1]. The third group, *compatibilists* argue, by contrast, that freedom and responsibility are compatible with the truth of causal determinism (and, according to the vast majority of them, also with most kinds of indeterminism). Incompatibilism—i.e., libertarians and hard determinists—and compatibilism are traditionally conceived as being mutually exhaustive positions.

The discussion between the parties has traditionally focused on the possibility of alternate possibilities. The challenge, as most of the authors took it, is, very roughly put, the following: if causal determinism is true, then (*ex hypothesi*) facts of the past, in conjunction with the laws of nature, entail every truth about the future (McKenna 2009). According to incompatibilists from causal determinism it follows that no one ever has a choice about anything, including her future actions. The most straightforward way to argue for this

[1] The main difference between hard determinists and hard incompatibilists lies in their reasons for denying the existence of moral responsibility. While classical hard determinists accepts the truth of causal determinism and deny the existence of moral responsibility on this basis, hard incompatibilists add the further claim that no possible indeterminacy could do any better in grounding moral responsibility the way we ordinarily conceive it.

conclusion is to accept some version of Peter van Inwagen’s Consequence Argument, which can confidently be called the most powerful argument for incompatibilism. The argument, in its most informal and sloppy formulation, goes as follows:

> If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (van Inwagen 1983, p. 56)

May it seem obvious, the Consequence Argument has been subject to many thorough criticisms since its first formulation. However, in order to defend compatibilism one need not refute van Inwagen’s main contention, i.e., that in the “absolute sense” (cf. Watson 1998)—that is, holding fixed *all* features of the past and *all* the physical laws—we do not possess the freedom to do otherwise. Rather, classical compatibilists argue that incompatibilists misconstrue the nature of the ability to choose (at least in the sense relevant to free will and moral responsibility). Although in the absolute sense we are not free to do otherwise, we still possess the capacity of choice in virtue of our actions being counterfactually dependent on our choices. Conditional accounts of freedom and responsibility claim that even if determinism is true, it is still the case that we could have done otherwise if we had chosen to do so. (Admittedly, choosing to do otherwise would require that either the past or the laws of nature were different from how they actually are. However, since most philosophers agree that these are contingent facts of the world, this assumption poses no further difficulties.) Classical compatibilists from Hobbes to Moore to contemporary authors (e.g., Fara, Smith, Vihvelin, whom Randolph Clark (2009) labels as the “new dispositionalists”) have put forward several arguments for the conclusion that conditional freedom is all we have in mind when talking about choice and alternate possibilities. Unsurprisingly, so-called leeway (c.f. Pereboom

2008) incompatibilists are not the least impressed by these lines of thoughts and maintain that freedom requires genuine choices in the absolute sense.

Incompatibilism, however, might have different roots. As opposed to *leeway incompatibilists* who claim that determinism is incompatible with free will and moral responsibility in virtue of depriving the agent of alternate possibilities, *source incompatibilists* argue that in a deterministic world no one can cause and control their actions the way required for free and responsible agency. Source incompatibilists typically claim that people in a determined universe lack the appropriate *authorship* over their own choices and actions; they cannot function as self-determining agents. Explications of the relevant notions of sourcehood or authorship vary from author to author. Here is a couple of examples: according to Robert Kane for one to be ultimately responsible for an action she has to be responsible for anything that is a sufficient reason for the action's occurring (see e.g. Kane 2002). According to Derk Pereboom determinism rules out moral responsibility because our actions result from a deterministic causal process that traces back to factors beyond the agent's (see e.g. Pereboom 1995, 2001). According to Galen Strawson (1994) we are truly responsible for our conduct only if we are responsible for who we are (mentally speaking)—however, since we cannot *ex nihilo* create ourselves, it is logically impossible to meet this criterion. Source incompatibilism, as examples clearly suggest, can have various intuitive roots and motivations, and accordingly source incompatibilist theories can take radically different stances on the issue of free will and determinism. While Robert Kane takes a libertarian position and maintains that metaphysically undetermined choices of a certain kind can and do establish ultimate responsibility, Pereboom holds that no scientifically credible form of physical indeterminism can yield us the required form of control. At the extreme, one can argue together with Strawson that true moral responsibility is logically impossible.

Putting all the subtleties aside, I think it is not unfair to say that until recently the concept of control primarily served the purposes of incompatibilists. However, since Harry Frankfurt's seminal paper "Moral Responsibility and Alternate Possibilities" (1969) has been published, this is far from being true. By providing a powerful argument for the conclusion that having the ability to do otherwise is not a necessary condition of moral responsibility, Frankfurt's paper has caused a considerable shift in the dynamics of the free will debate.

Frankfurt asks us to imagine the following scenario:

> Jones has resolved to shoot Smith. Black has learned of Jones's plan and wants Jones to shoot Smith. But Black would prefer that Jones shoot Smith on his own. However, concerned that Jones might waver in his resolve to shoot Smith, Black secretly arranges things so that, if Jones should show any sign at all that he will not shoot Smith (something Black has the resources to detect), Black will be able to manipulate Jones in such a way that Jones will shoot Smith. As things transpire, Jones follows through with his plans and shoots Smith for his own reasons. No one else in any way threatened or coerced Jones, offered Jones a bribe, or even suggested that he shoot Smith. Jones shot Smith under his own steam. Black never intervened. (McKenna 2009)

Frankfurt's point is straightforward: although Jones could not have done otherwise but to shoot Smith (had he changed his mind, Black would have interfered), intuitively he seems to be just as fully responsible for his action as he would have been if Black hadn't been around at all. Frankfurt draws the tentative conclusion that instead of arguing for the requirement of alternate possibilities we should formulate the responsibility-relevant kind of control by focusing on the actual processes which gave rise to the action.

So-called Frankfurt-type examples have attracted significant attention since the publishing of the paper and although quite a few, well-established doubts have been raised about their feasibility, their intuitive appeal gains many followers also nowadays. Frankfurt's example is usually considered as a powerful tool in defense of compatibilism. This is surely right: if Frankfurt-type examples are successful, they leave leeway incompatibilism ungrounded[2]. But it is important to note that Frankfurt-type examples constitute a threat not only for incompatibilists, but for anyone who defend the claim that moral responsibility stands or falls by our ability to make actions counterfactually dependent on our choices. Thus classical compatibilists—anyone who defends a conditional analysis of free will and responsibility—are challenged by Frankfurt to the very same extent as leeway incompatibilists are. Moreover, as John Fischer (2012a) argues, Frankfurt-type examples seem to be more resistant to objections coming from the compatibilist camp than to incompatibilist worries. Consequently, Frankfurt has not only munitioned compatibilists with a powerful argument; he also had a significant effect on forthcoming compatibilist theories.

For those compatibilists—sometimes called semicompatibilists—who build their strategy on the success of Frankfurt-type examples, the central question is this: what is it in the actual sequence of Jones's course of action which makes it an instance of responsible agency? Semicompabilists have to give a positive account without relying on alternate possibilities and counterfactual dependence. One dominant strategy to achieve this aim is to elaborate on the concept of *control*, which we normally exercise over our actions.

Frankfurt presents his own compatibilist solution in "Freedom if the Will and the Concept of Person" (1971), a paper no less influential than "Moral Responsibility and Alternate Possibilities". He distinguishes first-order—the desire to do something—and higher-order—the desire to have a desire—desires and argues that while freedom of action consists in doing what one wants, one possesses freedom of will if she wants what she wants to want (in Frankfurt's formulation: someone is a person if and only if she has higher-order volitions, that

[2] Source incompatibilists, by contrast, need not try to refute Frankfurt's contention.

is, higher-order desires which effectively influence one’s first-order desires). Although Frankfurt does not use the word “control”, it is relatively clear that freedom of the will is guaranteed by the influence of higher-order desires on lower-order desires. However, it is not quite clear whether according to Frankfurt freedom of the will is a necessary condition of responsibility; it rather seems that what is required for moral responsibility is mere conformity between one’s higher and lower order desires. The hierarchical account proposed by Frankfurt inspired many authors to further develop so-called “mesh theories”—represented by, among others, Gary Watson (1975), Susan Wolf (1990) and Hilary Bok (1998)—which claim that “a person is responsible for his behavior if there is an appropriate ‘fit’ between that behavior and various psychological elements of his or various features of the world” (Haji 2002, p. 203). For example, to take one of the most prominent accounts along these lines, Gary Watson, elaborating on a Platonic idea, distinguishes two sources of motivations, valuing and desiring, and claims that one acts freely when she does what she most values.

Mesh theories are not the only compatibilist rivals of classical compatibilism. Probably the most prominent recent attempt to define the responsibility-relevant concept of control has been presented by John Fischer and Mark Ravizza (1998). Fischer and Ravizza argue that freedom and responsibility are associated with the notion of guidance control as opposed to regulative control, i.e., the ability to choose between genuine alternatives. An agent exercises guidance control over her action if the mechanism which actually issues in her behavior is moderately responsive to reasons, where being moderately responsive to reasons means, in turn, that the mechanism is “’regularly’ receptive to reasons (some of which are moral), and at least weakly reactive to reasons” (Fischer & Ravizza 1998, p. 444).[3]

[3] According to Fischer and Ravizza guidance control also has a second, historical element: “taking responsibility” for the mechanism. For the sake of brevity here I will ignore this part of their account.

Peter Strawson is yet another author who inspired and provoked many compatibilist discussions in the last fifty years. In his seminal paper "Freedom and Resentment" he proposed that the practice of holding people responsible—and thus conditions of responsibility—should be understood in terms of those emotional and attitudinal responses with which we typically respond to other people's actions and omissions. Although most contemporary authors do not accept Strawson's strongest claim, i.e., that the inevitability and importance of reactive attitudes in human life make further metaphysical investigations unwarranted and "external", the nature and significance of so-called reactive attitudes such as guilt, gratitude or indignation have been much discussed in recent literature.

Overview

Attributionism is a fairly new type of theory of moral responsibility. The term *responsibility as attributability* was first introduced by Gary Watson in his thoughtful and inspiring paper, "Two Faces of Responsibility" (1996), in which he argued for a non-unified conception of responsibility. In a similar vein, in his influential book *What We Owe to Each Other* Thomas Thomas Scanlon distinguished two senses of responsibility, *substantive responsibility* and *responsibility as attributability* and provided a nuanced description and analysis of both concepts. Elaborating on the latter notion in a series of articles Angela Smith developed a unified account of moral responsibility, the "rational relations view" (see especially 2005, 2008a and 2012). Scanlon's and Smith's work, especially in recent years, got significant attention and, as a natural consequence of this, earned some committed opponents (see esp. Levy 2005, 2008, Levy & McKenna 2009, for various criticisms of Angela Smith's theory see e.g. McKenna 2008, Smith 2011, Shoemaker 2011) and followers (for instance, Pamela Hieronymi's work in the area of philosophy of mind and mental actions can rightly be called attributionist).

Aside from minor differences, Scanlon's concept of responsibility as attributability and Smith's rational relations view give the same model and justification of being responsible for something, understood as being open, "in principle, to moral appraisal—including moral praise and blame—on the basis of it (where nothing is implied about what that appraisal, if any, should be)" (Smith 2008, p. 370). According to Neil Levy's formulation, "on the attributionist account, I am responsible for my attitudes, and my acts and omissions insofar as they express my attitudes, in all cases in which my attributes express my identity as a practical agent. Attitudes are thus expressive of who I am if they belong to the class of *judgment-sensitive attitudes*" (Levy 2005). Judgment-sensitive attitudes include, among other things, beliefs, emotions and intentions, but also spontaneous reactions such as noticing something or caring about somebody. These attitudes are sensitive to agents' judgments insofar as in the case of an ideally rational person they come about if and only if the agent holds certain judgments.

Without going into any details we can already highlight some features of attributionist accounts, as presented by Scanlon and Smith, which can be legitimately called non-standard, compared to traditional theories of free will and moral responsibility. First, contrary to most theories which most often simply assume that we can be morally responsible only for our actions and voluntary omissions (and probably some consequences of these actions and omissions), for attributionist accounts the locus of investigation is the mental sphere—judgment-sensitive attitudes. Oddly enough, from attributionism it follows that we are only indirectly responsible for our actions—we are responsible for them as far as they are expressions of our judgment-sensitive attitudes (intentions, most notably). Second, as I will thoroughly discuss it in Chapter 8 ("Moral Criticism and Blame") attributionists provide an unusually weak concept of moral criticism, which they take to be the dominant form of holding each other responsible. And third, in both Scanlon's and Smith's works there are several attempts to undermine the importance of voluntariness, choice and consciousness in establishing the agent's responsibility. This is the point where issues about the responsibility-relevant kind of control becomes particularly pressing for attributionist accounts—in particular, it becomes unclear whether, according to attributionists, control plays any significant role in responsibility-attributions.

I think that these features of attributionism are deeply connected to each other and indicate some inner necessities which implicitly regulate theories of moral responsibility. In a nutshell, the connection is the following: the main motivation to elaborate on an attributionist account is to explain and justify some puzzling cases of responsibility: responsibility for attitudes and responsibility for involuntary omissions. That is, we regularly assess other people on the basis of their attitudes and involuntary omissions, even though it seems that we do not exercise the proper kind of control over these instances of our agency. It seems that our moral practices are inconsistent: on the one hand, we seem to be committed to an important moral principle, the Control Principle, which states that it is unfair to hold people responsible for things beyond their control. But, on the other hand, with our ordinary judgments of responsibility we frequently judge people on the basis of such things, over which, at least according some understanding of the concept, we do not exercise control.

There are several strategies to solve the tension between the Control Principle and our ordinary judgments of responsibility. Most typically, attributionists try to save ordinary judgments of responsibility by denying that control, as traditionally conceived by theories of free will and moral responsibility would be a necessary condition of moral responsibility. However, one cannot simply refute the Control Principle without explaining how the charge of unfairness can be replied. Thus, to throw this charge back, attributionists propose an unusually weak concept of blame and moral criticism, which, the argument goes, is not unfair even if directed toward attitudes or involuntary omissions.

In my thesis I wish to accomplish a dual aim. First, I would like to give a comprehensive and thorough analysis of attributionist accounts as presented primarily by Thomas Scanlon and Angela Smith. I will explore how they differ from apparently similar accounts, the strengths and weaknesses of their solutions to traditional problems of moral responsibility. I will raise several objections and investigate whether attributionist accounts have the resources to answer them. I do not volunteer to present a comprehensive defense of attributionism—I do not have an answer for every criticism. Rather, I would like to give a fair and sympathetic treatment of it which highlights why such an account might be preferable to its rivals. Attributionism has several major appeals and it is one of the explicit aims of the present discussion to present and enhance these.

Second, exploring attirbutionist accounts serves more general purposes. The analysis, as I indicated, will lead us to the discussion of the so-called Control Principle, i.e., the thesis that it is unfair to hold people responsible for things beyond their control. I would like to explore the problem emerging from the principle and give an abstract mapping of the possible solutions for it. This project requires us to investigate not only the concept of control—it raises fundamental and often neglected questions about the methodology of building up a theory of free will and responsibility and the division of labor between theories of responsibility and substantive normative ethical theories. By the end of the thesis hopefully I will achieve to give an abstract model of how theories of responsibility might respond to the problems emerging with the notion of control and to locate attributionist accounts within this framework. Also, this enterprise will shed light on some important methodological issues which often remain implicit in the relevant literature.

In Chapter 1 I will present the outlines of attributionist accounts and clarify their positions on a few crucial issues. Then, in Chapter 2, I will turn to the examination of the Control Principle and the inconsistent triad it generates. I will explore the available strategies to resolve the tension between the Control Principle and ordinary judgments of responsibility. I will lay special emphasis on the discussion and critical evaluation of *indirect* or *tracing* theories of responsibility for attitudes and negligent behavior, since they are the most promising rivals of attributionism. Then, in Chapter 3, 4 and 5 I return to the exploration of attributionist accounts. In Chapter 3 I explore the differences between hierarchical and attributionist accounts and further clarify the attributionist standpoint. The supposed differences can be nicely demonstrated by encountering the attributionist account of responsibility for carelessness and other forms of involuntary omissions—this is the task I undertake in Chapter 4. Finally, in Chapter 5 I turn to the attributionist treatment of responsibility for emotions as presented by Angela Smith.

Chapter 6 still focuses on the notion of control. In this chapter I present a conceptual analysis and argue that, according to our ordinary understanding, we do possess the ability to control our attitudes and involuntary omissions. This line of argument exemplifies a different strategy of dissolving the tension between the Control Principle and our ordinary practices than the one attributionists pursue, although there need not be any substantial disagreement between the two.

Chapter 7 introduces another major topic of the thesis: the relationship between *being* and *holding responsible* and the role of moral justification in theories of moral responsibility. Denying the inference from the unfairness of holding responsible to the negation of facts about being responsible is yet another strategy to eliminate the inconsistent triad generated by the Control Principle—moreover, apparently this is the strategy which Angela Smith follows.

In this chapter I defend the so-called normative interpretation, put forward as a general schema, from some objections (including Smith's) and discuss the methodological consequences following from its acceptance. According to the normative interpretation someone is morally responsible for something if and only if it is fair to hold her responsible on the basis of that thing—that is, facts about responsibility are defined by normative considerations regulating the fairness of responsibility-attribution. Accordingly, I will argue that any theory of responsibility has to define three variables: the scope of responsibility-attribution, the nature of the relevant responsibility-attributing practices and the substantive moral considerations about fairness which should be applied.

Chapter 8 undertakes the second task just mentioned: it explores the attributionist concept of blame and moral criticism and its relation to the charge that it is unfair to hold people responsible if certain conditions (e.g. exercising control) are not met. Attributionist authors, in particular Thomas Scanlon and Pamela Hieronymi, characterize blame and moral criticism by such milder practices which are supposedly immune to charges of unfairness. Thus, from a broader perspective the characterization of responsibility-attributing practices can serve as another way to bypass the problem generated by the Control Principle.

In Chapter 9 I take on the last issue which, as a consequence of the normative interpretation, should be explored: the concept of fairness. I analyze some allegedly independent notions of fairness as they arise in the debate between compatibilists and incompatibilists and try to show how they can be used to interpret the Control Principle. Here the remarks of Chapter 6 recur in the discussion: I will argue that different forms of lack of control invoke different notions of fairness and thus undermine responsibility for different reasons.

Finally, in the last chapter (Chapter 10) I will try to illuminate a perplexing issue: the division of labor between theories of responsibility and normative ethical theories. This topic enters the discussion at several points of the thesis. Consequently, this chapter is partly a summary and clarification of the points which I have previously made. Also, I will make some tentative claims about where attributionist accounts stand in this respect and what advantages and disadvantages this might bring about.

Chapter 1: Attributionism

1.1 Thomas Scanlon

Scanlon highlights the difference between the two senses of responsibility which he distinguishes with the following political example:

> It is said, for example, that there are two approaches to issues such as drug use, crime, and teenage pregnancy. One approach holds that these are the result of immoral actions for which individuals are responsible and properly criticized. The remedy is for them to stop behaving in these ways. The alternative approach, it is said, views these as problems that have social causes, and the remedy it recommends is to change the social conditions that produce people who will behave in these ways. Proponents of the first approach accuse proponents of the second of denying that individuals are responsible for their conduct. But this debate rests on the mistaken assumption that taking individuals to be responsible for their conduct in the sense of being open to moral criticism for it requires one also to say that they are responsible for its results in the substantive sense, that is to say, that they are not entitled to any assistance in dealing with these problems. (Scanlon 1998, p. 293)

Being morally responsible in the substantive sense entails judgments about "what we owe to each other", the main theme of Scanlon's book. Judgments of substantive responsibility are about what burdens and benefits we should take in social cooperation, how duties and entitlements are distributed among people. Although Scanlon does not formulate explicitly what "being substantively responsible" means, we could give roughly the following definition: one is substantively responsible for some action, consequence or state of affairs, if and only if according to those principles which no one with similar motivations would reasonably reject, she can have no claim for compensation or apology on the basis of it. Substantively responsibility need not be grounded in blameworthy action or conduct. On the one hand, if the appropriate conditions obtain, i.e., duties are properly distributed and fulfilled according to the right principles, responsibility evolves without any typical exercise of agency such as choice or decision. On the other hand, as the example above shows, even if we can legitimately criticize someone for her faulty conduct, without the appropriate conditions obtaining, the question of further requirements or duties remains open. The opportunity of choice becomes significant in this respect: since we attach both instrumental and non-instrumental (in Scanlon's classification: expressive and symbolic) values to it, those principles which no one would reasonably reject and which determine our duties must involve the opportunity of choosing among alternatives at least in certain cases. This is not to say, following an appealing tradition, that substantive responsibility requires voluntary control or decision) and thus the ability to do otherwise. We have independent reasons for valuing choice and these reasons must be manifested in the principles on which Scanlon's contractualism rests.

Responsibility as attributability, by contrast, concerns the assessment of the quality of one's self-governance. When we talk about responsibility in this sense, our concern is whether the given attitude (belief, emotion, intention) is properly attributable to the agent, that is, whether it is a proper basis for evaluating the agent. In this sense what we assess are the judgments the agent made about her own reasons, whether she gave proper weight to the proper considerations, whether she was able to see the right considerations at all. However, the target of responsibility-attribution is not the judgment itself, but those judgment-sensitive attitudes, which we take to reflect or express the judgments the agent holds. So the Scanlonian model looks roughly like this:

reasons → judgments about reasons → attitudes (→ actions)

Reasons are constant, given by both objective (facts about the situation, principles applying to it) and subjective (the agent's aims and preferences) factors. The region where rational self-governance takes place is the formation and reconsideration of judgments about the reasons the agent has. This doesn't consist in mere balancing of the weight of reasons, but is also the point where reasons might be altogether given up, if they confront with the agent's values and commitments. And finally, these judgments will determine which intention or belief the agent will form, or which emotion she will experience. The formation of judgments need not be either voluntary or conscious: what assure us about their existence are those judgment-sensitive attitudes which express them. And because these attitudes, as opposed to such things as, for instance, our height or eye color, reflect our judgments, we are morally responsible for them, "that is to say, we can in principle be called on to defend these attitudes with reasons and to modify them if an appropriate defense cannot be provided" (Scanlon 1998, p. 272).

1.2 Angela Smith

Angela Smith summarizes the core idea of the rational relations view as following:

> The view that I am putting forward takes as its starting point the idea that some of our mental states are linked to particular judgments in such a way that, if one sincerely holds a particular evaluative judgment, then the mental state in question should (or should not) occur. The »should« in question here is the should of rationality and, therefore, marks a normative ideal which our actual attitudes may not always meet (…). To take a simple example of the connection I have in mind: if I sincerely judge that there is nothing dangerous or threatening about spiders, I should not be fearful of them. The emotion of fear is conceptually linked to the judgment that the thing feared is in some way dangerous or threatening; therefore, my judgment that spiders are not in any way dangerous or threatening rationally entails that I should not be fearful of them. (Smith 2005, p. 253)

So, in comparison with Scanlon, the following model can be drawn:

evaluative judgments → attitudes (→ actions)

The only major difference compared to Scanlon's account which Angela Smith introduces is that instead of judgments about reasons she talks about evaluative judgments. However, even this modification seems to have minor significance, since both Scanlon and Smith emphasizes the dispositional character of these judgments and attitudes: while according to Scanlon "*Having* a judgment-sensitive attitude involves a complicated set of dispositions to think and react in specified ways" (Scanlon 1998, p. 21), Smith defines evaluative judgments as "continuing and relatively stable dispositions to respond in particular ways to particular situations" (Smith 2005, p. 251, fn. 27).

1.3 The Priority of Attitudes

As we could see, both Scanlon and Smith regard responsibility as attributability as primarily a matter of assessing one's attitudes as opposed to her actions. This shift of attention is both interesting and peculiar, since the main locus of investigation in the literature on moral responsibility has traditionally been responsibility for performing an action. This tradition can be explained in two ways. First, since even nowadays issues of moral responsibility arise almost exclusively within the framework of the free will debate, responsibility for actions has a prominent role in the discussion. This emphasis is natural, since one way of articulating the incompatibilist worry is that if determinism is true, then our experience of "can do otherwise", usually accompanying the performance of intentional actions, is illusiory. The phenomenology of performing an action seems to differ significantly from the formation of most mental states (mental images might be exceptions) and provides the *par excellence* example of doing something *at will*. Thus, as long as we assume that without this kind of freedom (which is experienced when we perform an intentional action) moral responsibility is impossible (and that appears to be a common incompatibilist position), exploring other kinds of self-governance or rational control characteristic of e.g. the formation of judgments seems to be unmotivated.

Second, at least from a moral point of view actions seem obviously more important than attitudes. Only actions and omissions and not attitudes can cause harm and injure others; we generally fulfill our duties and obligations by acting in certain ways (or by refraining from action). If morality has anything to do with other people, then it is mysterious how (even unexpressed) attitudes can gain priority over actions, no matter how sensitive they are to the agent's judgments.

Although these reasons make it justified in general to lay special emphasis on responsibility for action, this divergence from standard theories can be adequately explained within the attributionist framework. First, attributionist theorists are notoriously reticent when it comes to the metaphysical aspects of their accounts—Smith does not say a word about these issues. Scanlon, by contrast, at least briefly considers them and takes an explicitly compatibilist stand when arguing that responsibility as attributability is compatible with the truth of determinism and, more importantly, with a weaker claim which he calls the Causal Thesis, i.e., "that all of our actions have antecedent causes to which they are linked by causal laws of the kind that govern other events in the universe, whether these laws are deterministic or merely probabilistic" (Scanlon 1998, p. 250):

> These explanations of how various conditions can undermine moral blame do not lead to the conclusion that blame is always inapplicable if determinism, or the Causal Thesis, is true. The mere truth of those theses would not imply that our thoughts and actions lack the continuity and regularity required of rational creatures. It would not mean that we lack the capacity to respond to and assess reasons, nor would it entail the existence of conditions that always disrupt the connection between this process of assessment and our subsequent actions. So, even if one of these theses is true, it can still be correct to say that a particular action shows a person to have governed herself in a way that is morally deficient. (1998, p. 281)

This quote explains both Scanlon's unconcern about the genuineness of our experience of free action and the significance he attributes to attitudes despite their minor relevance in morality in general. First, Scanlon simply does not think that the relevant kind of freedom or control would be the one manifested in intentional actions. Note that this is a somewhat unusual stand even in the compatibilist camp: you don't have to be an incompatibilist to maintain that free choice and will have a central role in describing responsible agency. Scanlon, as I discussed it in the previous subsection, locates rational agency in the sphere of forming judgments about our reasons. There are several ways which make judgments of our reasons to be "up to us", i.e., we are able to revise and reconsider them, to understand and respect their reason-giving force (or the lack of it). Everything that takes place afterward (in the explanatory sense), seem to flow, according to Scanlon, inevitably from these judgments. Thus, as long as our attitudes and subsequent actions express these judgments, it does not matter if they come about involuntarily and uncontrollably—we still exercise our rational capacities.

It is important to emphasize that there is nothing peculiar in this action theoretical model. Since Davidson's groundbreaking work on this topic (see esp. Davidson 1963) most authors accept that our mental states *cause* our actions and this is a common assumption also among theorists in the free will debate. For instance, Galen Strawson's famous Basic Argument uses the following premise: "When one acts for a reason, what one does is a function of how one is, mentally speaking" (Strawson 1994, p. 6, see also Strawson 1986). Obviously, there are wide disagreements about whether this causation is deterministic or indeterministic, and exactly what kind of mental states are involved—these are the core individuating features of the theories. But Scanlon is surely not alone in thinking that how we choose and act crucially depends on our mental states. I take this to be the most fundamental root of source incompatibilist worries: even if we can establish some relative autonomy of choice or will, this success will only be apparent as long as our choice or will is decisively influenced by our beliefs and pro-attitudes, which we, in turn, do not control. Qualifications such as "true" (Strawson) or "ultimate" (Kane) responsibility usually refer to this further requirement: that it is not enough that our choice or will regulates our actions—we have to be able at least to some extent regulate those mental states which determine the choice or will itself. Authors, naturally, also disagree on the possibilities of such regulative control.

With this short interlude I wanted to highlight that it is not so much Scanlon's and Smith's action theoretical point of view which is non-standard, but the conclusions they draw upon it. In a way attributionist authors are admirably consistent. Their strategy can be interpreted as saying: once we realized that no choice or will is able to provide the sufficient degree of autonomy, freedom or self-governance (everyone can pick their favorite term), why should we further explore this region of agency? Attributionists suggest to take a step back and take a closer look on how these values can be realized in other spheres of mental agency. Obviously, attributionists are overtly optimistic in this respect: both Smith and Scanlon believe that the rational activity which takes place when we form, revise and hold judgments about reasons/evaluative judgments is enough to establish judgments of responsibility as attributability. There are some minor differences in emphasis, though: while Scanlon, similarly to traditional approaches, finds it crucial to retain the link between responsibility and some form of self-governance or "up-to-usness", Smith seems to be more eager to sever these ties once and for all. However, I am not quite sure that eventually these rhetorical niceties would boil down to substantive differences between them.

This almost exclusive focus on attitudes might still seem odd to many, but the considerations motivating it are relatively simple, if not straightly banal. The locus of responsibility as attributability is what Scanlon has later (2008) called the *significance* or *meaning* of an action or attitude, which is mostly determined by the agent's mental states such as her judgments and intentions. Of course these assessments of blaimworthy or blameworthy conduct can be made about, for instance, one's willingness to fulfill her obligations (manifested mostly in actions).

But what we are particularly interested in when holding someone responsible in the attributability sense is not the simple fact of norm violence, but what it tells us about the agent's commitments, values and endorsed reasons. As Watson poetically puts it:

> To adopt an end, to commit oneself to a conception of value in this way, is a way of taking responsibility. To stand for something is to take a stand, to be ready to stand up for, to defend, to affirm, to answer for. Hence one notion of responsibility—*responsibility as attributability*—belongs to the very notion of practical identity. (Watson 1996, p. 271)

Again, Scanlon might resist to characterize *his* notion of responsibility as attributability in these terms. But he would surely agree with Watson that actions have a meaning as far as they reflect and express the agent's judgments about reasons, which are doubtlessly constitutive to one's aims, values and commitments. We do not make moral assessments merely on the basis of the performance or omission of a certain action; the ultimate target of moral criticism is the agent herself, whose commitments, aims, endorsed reasons, etc. led to the action or omission. The same idea is present also in Smith's works. As McKenna bluntly puts it, when discussing her account: "what matters is how an agent judges, not how she acts. (…) [J]udgment is explanatorily basic for Smith, and actions are candidates for responsibility only to the extent that judgment is revealed in them" (McKenna 2008, p. 30).

Also, as we could see in the discussion of substantive responsibility Scanlon is well aware (and the same can be said about Watson) that responsibility has a much more diverse function than providing basis for negative or positive evaluation. Responsibility of attributability, and this is an important starting point of the forthcoming discussion, is only one form of responsibility-attribution which does not wish to explain such practices as, for instance, compensation or punishment. Thus it would be unfair to accuse Scanlon of elaborating on a peripheral phenomenon at the expense of more central and "serious" moral questions.

Smith's rational relations view seems more problematic in this respect, since Smith aims to give a unified account of the condition of moral responsibility "in the most basic sense" (Smith 2005, p. 237). However, this latter qualification can have a further meaning, i.e., that there can be other, "less basic" senses of responsibility, which, in turn, assume responsibility in the attributability sense, that is, that the given action or attitude is properly attributable to the agent. This idea has some plausibility: as we could see in the initial example of Scanlon, one can be morally responsible in the attributability sense, while not being substantively responsible for an action or outcome. But this relationship seems to be asymmetrical, since it would be rather difficult (if not impossible) to find an example for someone being substantively responsible while her action is not properly attributable for her. If that is true, then we can conclude that responsibility as attributability is not only one sense of responsibility, but that it is the weakest sense of it, which every stronger sense presupposes. Smith, however, does not elaborate on this issue, so from now on I will assume, in accordance with the most coherent interpretation of her articles, that she takes attributability to be the central notion of moral responsibility.

1.4 Control

Finally, it is of the utmost importance to clarify how attributionist accounts relate to the notion of control. Of course, there are some minor differences also in this respect. Watson is the most explicit when he says:

> [Real self views] are prompted by a concern with agency or attributability, rather than with control and accountability. (…) [I]ssues of control are subsidiary to issues of attributability. Control bears on attributability only so far as its absence that the conduct was not attributable to the agent. (Watson 1996, p. 271–272)

Watson talks about control in the same sense as the one we discussed when we explored source incompatibilists—that is, he discusses control which we exercise over our desires, ends, etc. which, in turn, determine our conduct. Also, this is the kind of control which he takes to be relevant to the Control Principle: "we can't be rightly blamed unless we have control over the causes of our conduct" (p. 269).

Smith, by contrast, seems to associate control with voluntariness. However, her characterization of the view which she labels "voluntarism" is rather ambiguous. McKenna is right to point out that

> In characterizing voluntarism, Smith sometimes speaks of what an agent has *voluntary control* over. At other points, she characterizes the view as restricted to what an agent has *chosen*, and still, elsewhere, to what an agent has *voluntarily chosen*. At one point she characterizes the view in terms of *deliberate choices*. Finally, in a footnote, she suggests that voluntarism is the view that responsibility is restricted to whatever an agent has *direct control* over. Depending upon one's theory of action, Smith's varying descriptions might capture very different views. (McKenna 2008, p. 30)

Although it would have been desirable to give a more careful description of her rivals, it is clear that Smith rejects any will, choice or deliberation-based account[4]. She puts forward a similar idea about voluntariness as the one Watson expressed with regard to control:

> In some cases (e.g., in determining a person's responsibility for a bodily movement), it may make sense to ask whether the agent has *voluntarily chosen* the thing in question, because that will determine whether that thing can reasonably be taken to express her judgments. But in other cases (e.g., in determining a person's responsibility for certain omissions, or for her desires, emotions, and beliefs), there

[4] As a matter of fact this imprecision on Smith's part might well be intentional. In her earlier article she writes: "What I have called the volitional view of responsibility is actually better understood as a cluster of distinct views which share a common assumption, namely, that choice, decision, or susceptibility to voluntary control is a necessary condition of responsibility" (Smith 2005, p. 238).

> does not seem to be any need to appeal to the agent’s choices, because the thing in question can be seen as directly expressive of her judgments and normative commitments. (Smith 2008, p. 368)

As I previously mentioned, Scanlon’s case is apparently different, since he lays special emphasis on the claim that judgment-sensitive attitudes are up to us. However, he is just as firm in stating that these attitudes might well arise in us without conscious choice or decision (Scanlon 1998, p. 22). He also points it out that our potential for self-governance does not come from the fact that our judgments about reasons sometimes make it permissible to choose between alternatives (ibid.). Moreover—and in this respect he interestingly diverges from the Watsonian distinction between attributability and accountability—he argues for a similar conclusion in the case of substantive responsibility, when he denies that the agent’s conscious choice between alternatives would leave all the possible burdens of the choice on her (this is the thesis which he calls the Forfeiture View, see pp. 258–260).

I find it pointless to further explore the notions of control which attributionist accounts refute. Putting aside the otherwise interesting divergences between the theories we can confidently attribute two important theses to them, both of which clearly distinguish them from the vast majority of theories of free will and moral responsibility. First, they all agree that, at least as far as responsibility as attributability is concerned, all traditional formulations of control (deliberation, choice, will, voluntariness, etc.) manifested exclusively in intentional actions are mostly irrelevant to judgments of responsibility. That it, contrary to what seems to be the common view, none of these are *necessary* criteria of moral responsibility—although most probably they constitute sufficient conditions of being morally responsible, in virtue of indirectly expressing one’s judgments (or aims or commitments, in Watson’s terms). Second, as a consequence of the previous point, the relevant agential activity should be found in the mental sphere: this is the formation and revision of judgments about reasons/evaluative judgments, which are expressed in judgment-sensitive attitudes. These instances of human agency are directly attributable to the agent, and consequently these are the things for which she is directly responsible.

Chapter 2: The Control Principle and Problematic Cases

The notion of control is one of the most often used concepts in the literature of free will and moral responsibility. As I mentioned in the introductory chapter, traditionally control was a useful tool in the hand of indeterminists, who argued that if any prior state of the universe, together with the laws of nature determines future states, then everything we do is beyond our control. Frankfurt-type examples, however, caused a shift in the debate between compatibilists and incompatibilists and various compatibilist notions of control have been developed.

It is already worth mentioning here that the concepts of control thus provided show a great variety: it is not only that they rely on different notions – such as reason, consciousness or choice –, but they differ also in their extensions: in some cases they do not agree on whether a certain action or omission is under the agent’s control. Here I would only like to illustrate my point by citing some influential (although no way comprehensive) formulations of the concept of control:

> To have control over one's actions, according to this picture, is to perform those actions intentionally, while possessing the relevant sorts of normative competence: the general ability to grasp moral requirements and to govern one's conduct by the light of them. (Wallace 1998, p. 86)
>
> To say that something is (directly) within my voluntary control is to say that I would do it (right away) if and only if I (fully) tried or chose or meant to do so, and hence that if I did it I would do it because I tried or chose or meant to do it, and in that sense voluntarily. (Adams 1985, p. 8)
>
> We exercise control over our actions only when consciousness has played a direct or indirect role in their production. (Levy 2007, p. 213)

> On our view, guidance control should be understood in terms of two elements: the agent's "ownership" of the mechanism that actually issues in the relevant behavior, and the "reasons-responsiveness" of that mechanism. (Fischer & Ravizza 1998, p. 241)

This heteronomy should come as no surprise and raises no philosophical problems in itself. Surely, competing theories of freedom and responsibility define the relevant notions differently and thus set different conditions of responsibility. However, an important thing should be noted: any informative account of control (or having control over something) has to be developed independently of issues of moral responsibility. The initial, intuitive idea behind introducing the concept of control is that people are morally responsible for some of their actions and omissions *because* they control them[5]. But for this very reason any such conception of control which simply picks out intuitively responsible agency without explaining how these are connected to control-exercising or its lack thereof, is explanatory superfluous. I do not want to claim that no account accomplishes this task, but I find that issues of control and considerations about responsibility are often conflated.

But there is a still prior question which calls for answer: why do we need the concept of control in the first place? As a matter of fact, it is difficult to find any compelling reason either on the side of the incompatibilists or the compatibilists. Just as powerful incompatibilist arguments, such as van Inwagen's consequence argument, can be easily put forward without relying on the somewhat obscure notion of control, nothing in Frankfurt-type examples compels us to explain their morals by the help of it.

I find that the insistence on using the concept of control can be adequately explained only by the proper recognition of an often neglected moral principle, i.e., that it is unfair to hold

[5] Maybe it would be more accurate to state it the other way round: the reason why we exempt or excuse agents from responsibility is that they did not exercise control. The latter formulation leaves place for those accounts according to which control is a necessary but not sufficient condition of moral responsibility.

people responsible for such things which are beyond their control. As George Sher (2005, 2009) rightly notes it, this principle—which he labels as Kantian, while Dana Nelkin (2008) calls the Control Principle[6]—is embarrassingly rarely discussed in the literature: "I think it is fair to say, (…) that although that principle is more often baldly asserted than carefully defended, it is even more often presupposed than explicitly stated" (Sher 2005, p. 180)

Although elsewhere Sher (2001) emphasizes that the principle seems to have a special "bedrock status", which makes difficult either to rebut or to defend it, we have several reasons to scrutinize it with the greatest precaution. Despite its strong appeal, apparent violations of the Control Principle are extremely common in our everyday judgments of responsibility and appear in a wide range of cases. To make the discussion smoother I distinguish three types of cases, which apparently violates the Control Principle. I will call these, in a somewhat question-begging, but hopefully innocent fashion, problematic cases of responsibility.

The first category could be labeled as *moral luck cases*, following Bernard Williams (1981) and Thomas Nagel (1979), especially *resultant* and *circumstantial* luck[7]. Resultant luck arises when the moral assessment of the action depends on its consequences, although these consequences arise independently of the agent's choice, will or intention. Arguably we assess differently a negligent driver who forgot to check the conditions of the brake, if consequently he hits a child than if he hadn't made any harm, because there was no one on the road. But it was beyond his control to determine the child's running onto the road. Circumstances,

[6] There are minor differences between their formulations, though. My formulation is almost identical to Sher's except that I use „things" instead of „wrongful actions and omissions". This modification is purposeful, since responsibility for attitudes (as opposed to actions and omissions) is one of the main topics I will discuss with regard to its relation to the Control Principle. Nelkin, by contrast, does not talk about fairness at all. Given the context, however, the 'cannot' included in her formulation can be understood only in a deontic fashion. About why this moral prohibition should be interpreted in terms of fairness, see the next chapter.

[7] Here I neglect the other two kinds of luck, the causal and the constitutive. The former is just another name of the dominant hard incompatibilist position, which claims that, under any coherent description of causation, our character and conduct are beyond our control (see the introductory chapter). The latter, roughly speaking, concerns responsibility for our traits and dispositions. I will not elaborate on this issue in particular, but most of what I will say about responsibility for attitudes applies also to character traits.

relevant to circumstantial luck involve all those social, political and other external conditions which influence the situations in which we act, the moral conflicts and dilemmas we face. Nagel's example is the Nazi collaborator, who, had he moved to Argentina in the '20's, wouldn't have faced the morally challenging choices of Nazi Germany and wouldn't have committed dreadful acts. However, now we condemn him for what he did, by contrast to his *alter ego*, who leads a completely normal life in Argentina. Yet, he evidently couldn't control the political atmosphere of his home country and the morally troublesome situations following from it.

The second category includes attitudes, broadly understood: *emotions*, *beliefs*, *intentions*, *desires*, *character traits*. Examples of moral criticisms based on people's attitudes, such as feeling envy for a friend's success, holding racist beliefs or simply being ungenerous or dishonest must be familiar to most of us. However, as virtually any author exploring these topics (see e.g. Adams 1985, Sher 2001, McCormick 2001, Goldie 2004, Owens 2000, Ben-Ze'ev 1997 & 2000) points it out, our control over attitudes seems to be either imperfect or entirely absent.

The third category collects instances of *involuntary omissions*: negligence, culpable ignorance, carelessness, forgetfulness, absentmindedness, etc. By involuntary here I simply mean that the agent did not *choose* to refrain from performing a morally relevant action, but the omission is explained by lack of attention, the absence of due care or forgetfulness. Since in these cases the agent is not even *aware* of her omission, it is plausible to ask how she could control it. For the sake of brevity I will refer to this category as cases of *negligence*.

So apparently we have a yet unanalyzed, but definitely appealing moral principle and a wide range of cases which seem to violate this principle. Consequently, for our practices to be morally justified, one of the following premises has to be given up:

P1: It is unfair to hold people responsible for things over which they do not exercise control. (Control Principle)

P2: In problematic cases of responsibility the agents do not exercise control.

P3: In problematic cases the agents are morally responsible.

It is important to note that puzzles about control and responsibility arise on different levels of generality. Whereas the classical incompatibilist worry of our character and conduct being beyond our control threatens all instances of human agency, resultant and circumstantial luck cases might be eliminated without undermining the whole practice of holding each other morally responsible (although this claim is debatable). And the same can be said with greater confidence about responsibility for attitudes and negligent behavior—although giving up assessing and blaming people on the basis of these would require major revisions in our moral practice, without additional arguments we have no reason to think that these revisions would also endanger the possibility of responsibility-attribution for actions, voluntary omissions and consequences. As George Sher puts it at the beginning of his discussion of responsibility for negligent behavior:

> Unlike Nagel, I am interested not in the generic and necessary ways in which agents lack control over what they do but rather in certain merely contingent factors that also seem inimical to control. Because these contingent factors are not universally present, this threat to responsibility is not universal either, but because the factors are often present when we hold agents responsible, the threat remains significant and far-reaching. (Sher 2006, p. 286)

Given these important differences it becomes more understandable why there isn’t much communication between the different parties. On the one hand, authors contributing to the traditional free will debate are rarely interested in those “merely contingent” factors which threaten responsibility quite independently of any metaphysical consideration. Authors like

Sher, on the other hand, do not have much to say about the metaphysical issues. This might seem odd to some since, as I just told, metaphysical worries arise on a universal level and in this sense they seem to be prior to local, contingent problems. However, there is nothing peculiar in this methodology. According to my understanding those authors who focus primarily on particular puzzles about responsibility implicitly assume that *whichever position you take* in the free will debate, it will not solve the puzzles which arise at a lower level. That is, even if we assume that we have a satisfactory answer for the metaphysical worries, this answer will not automatically help in handling problematic cases. Pre-theoretically we tend to assume that we are free and responsible agents and find this idea quite unproblematic; still we find moral luck cases, responsibility for attitudes and negligent behavior troubling. This suggests that the different levels can (although certainly need not) be separately discussed.

Before we proceed, it is worthwhile to summarize the available strategies to dissolve the tension. Although the Control Principle itself has indeed got little attention in the literature, there is a vast amount of philosophical work which attempts to conciliate moral luck cases with our normative commitments. And there is also a rapidly growing body of literature which aims to explain and justify responsibility for attitudes and negligent behavior.

First, we can deny the relevance of the Control Principle (P1) altogether, thus providing a uniform solution for all the cases listed above. Put it as flatly as I did, I cannot see how this task can be accomplished without a highly revisionary analysis of the principle. As Thomas Nagel rightly points out,

> When we undermine moral assessment by considering new ways in which control is absent, we are not just discovering what would follow given the general hypothesis, but are actually being persuaded that in itself the absence of control is relevant in these cases too. (Nagel 1979, pp. 26–27.)

Although I do not accept Nagel's final diagnosis—i.e., that moral luck cases create a genuine paradox, which cannot be resolved by rethinking the conditions of moral responsibility—denying that the Control Principle *seems to* have a relevant application in these cases would be theory-driven blindness. Luckily, real life defenders of this strategy are much more nuanced: most often they argue for an alternative condition of responsibility which is meant to reserve some of our intuitions motivating the Control Principle, while making room for problematic cases. This is the strategy which, for instance, Smith and Watson follow when they argue that control is relevant to responsibility only so far as it indicates that the agent's behavior expresses her evaluative judgments.

It is a different but related way of arguing along these lines to say that to achieve a better moral outlook it is morally preferable to abandon the Control Principle (see Walker 1991). Note that these types of arguments (would they succeed) need not deny neither the *prima facie* normative force of Control Principle, nor its clash with ordinary responsibility judgments—they rather claim that in various cases the Control Principle is defeated by another moral consideration.

It is a far more common strategy to deny the violation of the Control Principle by claiming that our practices do conform to the principle (refuting P3). In moral luck cases, as we have seen, problems arise when contrasting the moral assessments of two agents, who are identical in every controllable respect. Those who want to deny the existence of moral luck have to show that assessments in the two cases do not differ after all, or, more plausibly, that the differences can be explained by epistemic limitations (see e.g. Thomson 1993) and thus be mitigated. However, the two latter categories do not involve any similar comparison. Consequently, the only way to deny the principle violation would be to argue that we do not in fact hold people responsible for their attitudes and negligent behavior. Although at first glance this strategy does not seem to be any more fruitful than rejecting the Control Principle, some authors have made attempts to show that the evaluative judgments we form about someone's traits or attitudes do not imply responsibility-attribution (Levy 2005), while others argue for similar conclusions about negligence (King 2009). Alternatively, one can formulate her revisionism in a prescriptive manner: given that the Control Principle is valid, we *should* stop holding people responsible for things which are beyond their control. Interestingly, revisionist accounts of this kind often oscillate between the descriptive and the prescriptive interpretation.[8]

Finally, you can argue that under the most adequate interpretation of the notion of control problematic cases do not violate the Control Principle (denying P2), since we *do* exercise—direct or indirect—control in these cases. This strategy is obviously not available in resultant and circumstantial luck cases—there is no sensible way to argue that it is in any significant sense within our control to influence either the consequences of our actions or the external circumstances. However, in the case of responsibility for attitudes and negligent behavior this is currently the most typical strategy to dissolve the clash between the Control Principle and our actual practices. Since these accounts are the main rivals of attributionist theories, it is inevitable to explore them in details.

The more radical line of argument leads to the conclusion that we have *direct* control over our attitudes and/or negligent behavior. It might come as no surprise that these accounts are fairly rare. Given the notions usually associated with control—such as choice, voluntariness and consciousness—this claim sounds utterly counterintuitive and such accounts are usually criticized on this basis. (For an example, see Solomon 1973, who argues that we have exactly the same kind of control over our emotions as over our actions.)

[8] Hopefully my discussion of the normative interpretation will illuminate why this might be so.

However, direct control accounts need not come up with such extraordinary claims. They can instead redefine the concept of control so as to make it compatible with judgments of responsibility in the problematic cases. This is the strategy which, for instance, George Sher (2005) follows in handling cases of negligence:

> Instead of having to say that any act of whose wrongness an agent is unaware is necessarily beyond his control, we may be able to say that whether such an act is within his control depends on whether its wrongness is suitably related to whichever combination of conscious and nonconscious attitudes is properly his own. (Sher 2005, p. 296)

Although coming up with a revisionary conception of control has some advantages, it is obviously vulnerable to the criticism that although the revised conception might well be useful for some purposes, it is evidently not the *relevant notion* of control. In his response to Sher, Neil Levy (2008) puts forward this claim, when he writes:

> Suppose an agent satisfies Sher's condition for responsibility in the absence of consciousness; suppose, that is, that his failure to believe that he is acting wrongly is explained "by some subset of the other beliefs (desires, attitudes, etc.) that make him the person he is." In what sense, however, does the fact that he satisfies this condition make it the case that he exercises control over his action? The relevant control problem arises, recall, because we do not exercise control over anything of which we are unaware. Sher's condition explains why the agents in his cases are not aware of the wrongmaking features of their actions, but explaining why they are not aware of these features explains why they do not exercise relevant control: it does not restore it. Sher's solution leaves the relevant control problem untouched. (Levy 2008, p. 216)

Also, now we can see that this last strategy which Sher pursuits shows startling resemblance with the first one which I presented, i.e., arguing directly against the Control Principle. The inconsistency between the three premises comes about only if control is used in the same sense in P1 and P2. Both strategies are based on the assumption that we are guilty of equivocation and present a revisionary conception of control which does not exclude problematic cases.

Also, both strategies are vulnerable to the same kind of criticisms. On the one hand, those who insist that the Control Principle is incompatible with one or another problematic case can always claim that the revised conceptions of control are faithful to the Control Principle only in their wordings. Control requires, according to their understanding, more stringent conditions which, contrary to what our ordinary responsibility judgments suggest, often cannot be met. On the other hand, despite the revolutionary tone of some opponents of the principle, virtually no one denies that there is some truth in the Control Principle which should be appreciated[9]. For these reasons I take the difference between the two strategies to be merely terminological. This obviously does not imply, however, that there aren't differences in the degree to which authors take the Control Principle seriously.

Indirect or *tracing* theories are quite different and constitute the most numerous and popular camp when it comes to explaining responsibility for attitudes and negligence. A clear-cut example for the latter is Holly Smith's account of culpable ignorance (1983). Smith assumes that in all cases of culpable ignorance there is a sequence of actions: a so-called "benighting act", when the agent "fails to improve (or positively impairs) his cognitive position" (p. 547) followed by the "unwitting wrongful act". To take her central example: the doctor who, unbeknownst to him, caused severe eye damage for a premature infant, because he used unnecessarily high concentration of oxygen, is blameworthy for blinding the infant because at a prior time he omitted to read the latest issue of the medical journal, which published a study telling about these effects. According to Smith, the following things must be true in order to rightly blame the agent for committing the unwitting wrongful act (blinding the infant): (1)

[9] Eugene Schlossberger might be a counterexample.

the benighting act (not reading the journal) must be culpable, i.e., it has to be morally wrong and the agent has to be responsible for committing it and (2) the unwitting act must fell (known to the agent) “within the risk” of her benighting act, that is, the agent must be aware that with her culpable action or omission she runs the risk of committing the latter unwitting act.

Another example of so-called tracing theories, this time from the sphere of attitudes, can be found in Fischer & Ravizza (1998) who discuss moral responsibility for emotions. Fischer and Ravizza only sketch the outlines of an account which is the direct expansion of their theory of responsibility for actions, omissions and consequences, which is based on the concept of guidance control. Since earlier I briefly discussed their theory, here I won’t recap it. Its application to emotional reactions would look roughly like this:

> What is natural to require (…) is that we can *trace* back to *some* appropriate point in the agent’s past and find an exercise of guidance control that then results in the subsequent emotional reaction. The subsequent emotional reaction must be the result of guidance control at *some* suitable prior time, in order for the agent to be morally responsible for the emotional reaction. (Fischer & Ravizza 1998, p. 255)

It is worthwhile to highlight, on the basis of the accounts just presented, the common features of tracing theories. Tracing or indirect theories claim that we are indirectly responsible for an attitude or involuntary omission (or anything else, by the way), if it is a result or consequence of an earlier action or omission for which we are directly responsible. The core idea is that the control which we exercise when we perform free and responsible actions is transferred to some of the consequences of the action and thus the traditional connection between responsibility and control can be reestablished. For tracing theories responsibility for actions (and omissions) is basic and take responsibility for non-actions as derivative from it.

So far, so good. It would be hard to deny that there is something obviously appealing about tracing theories. First, they preserve the connection between control, voluntariness and responsibility. And second, they reflect an important intuition of ours: that exercising control over something means, at the very minimum, that *we can do something about it*. Arguably, we would think differently about our attitudes and involuntary omissions if we knew for sure that we do not have any means whatsoever to prevent them. Every time we hold someone responsible for such problematic cases we implicitly assume that there was something the agent could have done, even if only in principle, to avoid the wrongdoing.

But tracing theories have constant and notorious problems regarding the scope of responsibility-attribution. That is, it seems that indirect theories can explain only a small, if not negligible subset of those cases for which we ordinarily hold people responsible in the absence of voluntary control. To illustrate the typical shortcomings of tracing theories, take George Sher's often cited example, *Hot Dog*:

> Alessandra, a soccer mom, has gone to pick up her children at their elementary school. As usual, Alessandra is accompanied by the family's border collie, Bathsheba, who rides in the back of the van. The pickup has never taken long, so, although it is very hot, Alessandra leaves Sheba in the van while she goes to gather her children. This time, however, she is greeted by a tangled tale of misbehavior, ill-considered punishment, and administrative bungling which requires several hours of indignant sorting out. During that time, Sheba languishes, forgotten, in the locked car. When Alessandra and her children finally make it to the parking lot, they find Sheba unconscious from heat prostration. (Sher 2006, pp. 286–287.)

Most of us would agree that Alessandra is responsible and blameworthy for risking Bathsheba's life, although, in the traditional sense, she did not control her forgetfulness. If we are to explain Alessandra's responsibility by means of an indirect account, we have to trace Alessandra's responsibility back to a prior action or omission over which she had control.

The most fundamental problem, as Sher rightly points it out, is that we do not find any suitable candidate for this role. What should Alessandra have done in order to ensure that she wouldn't leave the dog in the car? Since the row at the school was unexpected, Alessandra could see no reason to break the daily routine which proved to be safe and comfortable for all parties.

Yet, literally speaking there would have been countless ways to prevent her forgetfulness. For instance, if she hadn't got so deeply irritated by the headmaster's tone, it surely would have come into her mind that Sheba was in the car. However, this obviously won't do, since getting irritated is clearly not something over which we have voluntary control. The prior event from which the agent's present responsibility is derived has to be an undisputable case of controlled, responsible agency—otherwise we cannot re-establish the connection between responsibility and control.

Finally, let's say that we find such a prior action or omission. Some would say that at the end of the day Alessandra's fault was to carry the dog with her instead of leaving her in the air-conditioned apartment. However, as I previously said, the happenings in the school were quite unexpected, so Alessandra could not reasonably foresee that by carrying Sheba she runs the risk of leaving her in the hot car for hours. But how could we hold her responsible for the consequences of her forgetfulness, if she couldn't possibly foresee that her prior voluntary actions and omissions would lead to such a terrible result.

When we generalize all these requirements, it turns out that cases of indirect responsibility are hard to come by. We need to find such a prior action or omission, where the agent's responsibility is undisputed and it was reasonably foreseeable (or, depending on the particular theory, actually foreseen) for her that this prior action or omission might result in the present wrongdoing. In the case of attitudes it is even more improbable to meet these requirements—

for instance, sometimes it is even hard to say what *kind* of action or omission the agent should have performed. What would have prevented me from being treacherous or feeling envy of my neighbors? But even when we can identify such precautionary steps, it is extremely improbable that they would be such voluntary actions or omissions, which led foreseeably (that is, for the agent at that prior time) to the present attitude.

These charges are raised quite frequently against indirect theories (see e.g. Adams 1985, Sher 2005, Vargas 2005) and are regularly refuted (with more or less success) by the theory's representatives (see e.g. Fischer & Tognazzini's reply to Vargas (2009)). I will not follow this debate any further, but would like to raise another objection which seems especially relevant in the present context.

Even if we put aside the problems concerning the scope of responsibility according to indirect theories, one might say that these accounts identify incorrectly the *target* of responsibility-attribution. That is, although it is true that when holding an agent responsible for something we implicitly assume that she could do something about that thing, this isn't *why* we hold her praiseworthy or blameworthy. The difference between the supposed *conditions* of responsibility and the *content* of responsibility-attribution is especially apparent in the case of attitudes. Let's say I criticize someone for being contemptuous of people with less fortunate social circumstances than her own. Arguably, this criticism would be morally inappropriate if the agent couldn't do anything whatsoever to get rid of her contemptuous attitude. But *what* I find blameworthy is not her past omission to cultivate better attitudes or something along this line, but the attitude itself and all those judgments and commitments which it presumes. The wrong-making features of her attitude are to be found *in the attitude itself*, not in some past exercise of control.

Obviously, one could say that these are two separate issues: when defining the conditions of responsibility we need not give an account of the sources of blameworthiness and praiseworthiness. I am not sure whether after careful examination this would turn out to be a viable response—but it is certainly unsatisfactory in the case of indirect theories. Since at the end of the day tracing accounts claim that any kind of responsibility which we indirectly bear for our attitudes and negligent behavior is derived from our prior, responsible actions and omissions, it is hard to see how we could be any more or less blameworthy for the former than the latter. As Smith points it out:

> On this view, to say the culpably ignorant agent is to blame for his unwitting act is to say nothing more or less than that he was culpable in performing the benighting act, that it gave rise to the unwitting act, and that he knew at the earlier time that he risked this outcome. But we knew this at the outset—indeed, this is just a description of what *makes* something a case of culpable ignorance. Your claim turns out to be much less bold than I thought it was, since after all, *you are not attributing any independent fault to the agent* [my emphasis] beyond his fault in culpably performing the benighting act. (Smith 1983, p. 566)

I find this an especially important objection against indirect theories. Once we thoroughly understand the spirit of these accounts, it would naturally follow from them that we can be blamed on the basis of our attitudes and negligent behavior only to the degree to which we are blameworthy in committing the past action or attitude for which we are directly responsible.

This latter claim, however, is boldly refuted by everyday experience. Even if we find a prior action or omission to which the present wrongdoing can be traced back, it is all too often the case that while the past omission is of relatively minor moral importance and would never come to our mind to actively blame the agent for it, it results in such moral deficiencies to which we are resolutely opposed. To illustrate this point here is a nice and accurate example by Adams:

> The morally imprudent voluntary omissions, for example, by which a person has failed to pay the price to extricate himself in time from a situation that has left him embittered, cynical about morality, and full of racist resentments, may be less gravely blameworthy than the at-itudes to which they have led. Indeed we might think them blameless, a successful gamble, if the sequel had not left the person so corrupted. These considerations confirm the intuitively plausible judgment that what we chiefly blame in the present immoral state of mind is not the imprudence of the previous voluntary omissions. (Adams 1985, p. 14)

In this chapter I argued that the concept of control plays a central role in discussions about responsibility because of an intuitively appealing but yet unanalyzed moral principle, the Control Principle, which states that it is unfair to hold people responsible for such things which are beyond our control. However, certain problematic cases seem to contradict to the principle: resultant and circumstantial luck, responsibility for attitudes and negligent behavior seem to admit to a wide range of cases where we judge the agent responsible despite of the fact that she did not exercise control over the given thing. I explored the available strategies to dissolve the clash between the Control Principle and ordinary judgments of responsibility, with a special emphasis on so-called tracing theories.

Now that we encountered the main rival of attributionist accounts, we are in a better position to recognize their shortcomings and appreciate their advantages. But before turning to their treatment of responsibility for attitudes and negligent behavior we first need to distinguish attributionism from another group of theories, which I previously labeled as *hierarchical accounts* or *mesh theories*. After then, in Chapter 4 we turn to the attributionist treatment of involuntary omissions, which is followed, in Chapter 5, the exploration of Angela Smith's account of responsibility for emotions. Since responsibility for attitudes encompasses several different mental states (beliefs, intentions, desires, etc.), it would be a monstrous job to present the attributionist solution for these cases one by one. Hopefully my discussion of

Smith’s theory with regard to emotions will illuminate the general structure of the attributionist solution and also provide some hints, how these accounts can be further improved.

Chapter 3: Attributionism and Hierarchical Accounts

In Chapter 1 I presented the outlines of attributionist theories and also discussed some of the major differences between Thomas Scanlon's, Angela Smith's and Gary Watson's accounts. One might find, however, Watson's role somewhat unclear in this context. In the introductory chapter I recognized him as one of the most prominent representatives of so-called hierarchical or mesh theories—theories, which, loosely speaking, identify the central condition of responsibility with some sort of fit between the agent's conduct and certain psychological elements of her. So, for instance, Frankfurt argues that one's will is free if her higher-order desires align with her first-order desires which, in turn, regulate her behavior, while according to Watson one is free if her behavior is in line with her valuational system. Moreover, the distinction he draws between attributability and accountability is part of the defense of his account against Susan Wolf's objections (1990). When Wolf and Watson talk about "real self views", they primarily mean some version of mesh theories. However, I previously assumed that attributionism is a distinct view which should be distinguished from hierarchical accounts. How do these claims come together?

Without reaching unnecessary philological depths in exploring the contemporary philosophical literature, it might be worthwhile to mention that in "Two Faces of Responsibility" (1996) Watson does not explicitly defend the account which he put forward in his "Free Agency" paper (1975). Rather, his aim is to present a somewhat unusual way of thinking about agency, responsibility and attributability and argue for its legacy. The distinction he draws might help defending his account from Wolf's objections, but it is intended to deliver more general morals. Thus, it might be pointless to require complete coherence between the two papers. This, however, does not answer the other, more substantial question: why shouldn't we regard attributionism as a subset of hierarchical accounts?

One could easily neglect these classificatory issues as pointless verbal disputes. However, the differences are quite real and worthy of investigation. First and foremost, I find that contrary to attributionists hierarchical accounts retained the traditional focus on actions. In Watson's case this is evident: he intends to account for free (as opposed to merely intentional) actions and his action theoretical apparatus is restricted to factors which motivates one to *act*.

Frankfurt's case is much less obvious, since he cautiously differentiate between freedom of action and freedom of the will, providing emphatically a theory of the latter. Moreover, in other works (1976, 1987) he clearly affirms responsibility for attitudes, most notably for desires. Still, his emphasis on desires is somewhat telling: at the end of the day mental states are significant as far as they regulate human conduct and the genuineness of desires has to be guaranteed in order to make our behavior truly "our own". I am aware that some might find these claims suspicious and ask (reasonably) for further evidences to underline them. Although here I cannot volunteer to take on a more detailed discussion of this topic, I take it as a further, implicit evidence for my view that Frankfurt remains silent about every other kinds of attitudes, whose role in regulating behavior is less obvious.

It would be unreasonable, though, to deny all that there is a lot in common in attributionist and hierarchical theories. Scanlon himself also recognizes the similarities with Harry Frankfurt's account, and discusses his key example in details:

> Frankfurt considers two agents who, it is supposed, are both addicted to a drug to a degree that makes it impossible for them to resist the temptation to take it. One of them, the "unwilling addict," objects to having this addiction (in Frankfurt's phrase, he does not want the desire to take the drug to be effective in determining his action); whereas the other, the "willing addict," prefers the life of addiction, wants to act on the desire to take the drug, and would do so whether or not he was able to resist. Frankfurt says that when the willing addict takes the drug he acts freely in the sense relevant to moral responsibility, despite his inability to act

> otherwise. The unwilling addict, on the other hand, is not free in this sense. The account I have offered supports this conclusion, provided that by "moral responsibility" we mean the attributability of the action to an agent. Since the action of the willing addict reflects his assessment of the relevant reasons, he acts freely in the sense required by this notion of responsibility. The unwilling addict, on the other hand, does not act freely in this sense. (Scanlon 1998, p. 291)

There are several points to be made here, which might get us closer to deeper understand the attributionist account which Scanlon proposes. First, introducing the Frankfurtian example this way clearly suggests that, similarly to Frankfurt's original theory, for an action and attitude to reflect the agent's judgment of reason there need not be any kind of causal connection between the judgment and the action/attitude. On one plausible interpretation of Frankfurt's influential article, what is needed for moral responsibility is not free will in the sense of our second-order desires causally controlling the first-order ones, but only conformity between higher and lower-order desires. Similarly, according to Scanlon's account, for an action to be reflective of one's judgments of reasons (and thus be one for which the agent is morally responsible), we only need to establish that the agent actually holds that judgment of reasons which we can reasonable infer from his behavior.

But these results are problematic. The basic trouble is that it is puzzling what a judgment-sensitive attitude might be, if those who are "addicted to a drug to a degree that makes it impossible for them to resist the temptation to take it", can also possess it. The most natural way to interpret judgment-sensitive attitudes is to make the following counterfactual claim: "If she had a different judgment of reasons, she would have a different attitude". But this condition of judgment-sensitivity clearly does not apply to Frankfurtian addicts: no matter what judgments they actually hold, they will take the drug. And if that's true, then neither of the addicts can be held responsible for taking a drug.[10]

But there is another way to solve the tension consistently with Scanlon's and Smith's account. Smith's defines the conditions of moral responsibility in the following vain: "In order for a creature to be responsible for an attitude, on the rational relations view, it must be the kind of state that is open, in principle, to revision or modification through that creature's own processes of rational reflection" (Smith 2005, p. 256). This "in principle" closure makes it possible to give a different interpretation of Scanlon's and Smith's claims: probably the condition of judgment-sensitivity should be applied not to particular attitude-tokens, but to kinds of attitudes. Following this line of argument Scanlon could say that the action-type of taking drugs is in principle judgment-sensitive and thus the agent is open to moral criticism on

[10] John Fischer (2012a) recently argued that Scanlon's concept of judgment-sensitivity faces the very same difficulties as conditional accounts in general. That is, as I briefly mentioned in the introductory chapter, Frankfurt-type examples show that no conditional analysis can adequately capture the control condition of moral responsibility, since we can always imagine a counterfactual intervener, who blocks counterfactual dependence without thereby undermining moral responsibility. Here is Fischer's Frankfurt case in Scanlonian terms:

> In a suitably revised Frankfurt case, the agent (say Jones) makes a judgment as to what is best based on his own reasons, and he is in no way impaired or interfered with. Further, this judgment issues in an appropriate attitude and also subsequent behavior. Intuitively, Jones acts freely and is morally responsible. But, given the presence of Black, it is true that had Jones judged differently, Black would have swung into action and induced the very same attitude and behavior that occur in the actual sequence of events. Thus, Jones's attitude is not judgmentsensitive (where judgment-sensitivity is defined in terms of subjunctive conditionals of the sort employed in the conditional analysis of freedom), and his behavior does not flow from a judgment-sensitive attitude. (Fischer 2012a, p. 146)

I find Fischer's argument convincing and troubling for the attributionists (I assume that the same objection holds in the case of Smith's account). Obviously, there are several arguments questioning the effectiveness of Frankfurt-type examples, which attributionists might find suitable for their purposes. But at this point they have to admit that they cannot afford the luxury of simply ignoring the debate about alternate possibilities anymore. I find it a particularly appealing feature of attributionist accounts that they bypass virtually all the traditional metaphysical issues of the free will debate. Once they have to take a stand on these matters, we loose an important reason for embracing attributionism as opposed to other compatibilist accounts.

If attributionists are to stay neutral on metaphysical questions raised by Frankfurt cases, then they have to give a non-conditional analysis of judgment-sensitivity. However, I cannot imagine even the outlines of such a solution.

the basis of it. However, it is clear that this strategy is just as incapable of discriminating between the willing and the unwilling addict as the former one. If we are morally responsible for all our in principle judgment-sensitive attitudes, then there is no difference between those whose attitudes conforms to their consciously held judgments and those whose attitudes do not. Thus the distinction between the two addicts with respect to moral responsibility still cannot be established.

This discussion suggests that hierarchical compatibilist accounts (or at least the Frankfurtian version of it) have obvious advantages over attributionist theories. However, I do not think that Frankfurt's two addicts would be such a paradigmatic and clean-cut example of the difference between responsible and non-responsible agency that any feasible theory of responsibility has to lead to the same results as Frankfurt's account. But the decision whether to apply the condition of judgment-sensitivity to attitude-types or tokens still has to be made. I find that both options have both independent and interrelated difficulties, which are worthy of attention.

The token version of attributionism faces a fundamental epistemological problem: how can I know that the agent actually holds the judgment which her attitude reflects? In the case of "real" judgments, i.e., consciously held beliefs, the situation is easy: the agent either have a judgment supporting the attitude or one which in some sense contradicts to it (as in the case of recalcitrant emotions and beliefs). In the former case she is responsible, while in the latter she is not. But what should we do if there is no conscious judgment around? Here comes what Angela Smith calls rational inference from the agent's attitude to her underlying judgment. But what makes such an inference valid? I think the only reason which allows for such inferences is that we know about the given attitude-token that it belongs to the kinds of attitudes which are in principle judgment-sensitive. It seems then that the token-version of attributionism is unstable and every time when the attitude is not supported by a conscious judgment on the agent's part (which is fairly common), it collapses back into the type-version of it.

So why not accept the type-version? Just to remind, this would amount to say that people are morally responsible for all those attitudes, which belong to a judgment-sensitive attitude-type. This strategy, which Smith defends explicitly while Scanlon seems to be somewhat hesitant, is more promising. But in this case everything depends on how we identify and classify these attitude and action-types. In the case of attitudes the situation is a bit easier: given what we know about these attitudes, we have good reasons to agree with Smith that, for instance, random thoughts are much less sensitive, if they are at all, to the agent's judgments than beliefs and emotions. This can serve as a basis of justification for not holding people responsible for their random thoughts while we do hold them responsible for many of their beliefs and emotions. But can we do similar distinctions in the case of actions, which are, at least in the case of intentional actions, obviously part of responsible agency by virtue of expressing judgment-sensitive attitudes? As we have seen in Frankfurt's example here the situation is more complex. While an action described as "taking a drug" satisfies the condition of judgment-sensitivity, "acting on an irresistible desire", if there is such, surely does not. Similarly, although washing one's hand is usually part of responsible agency, doing it 147 times a day due to suffering in obsessive-compulsive disorder is obviously not. Of course the reason why we exempt agents from responsibility in these cases has a lot to do with in principle judgment-sensitivity. However it won't do for the attributionists to simply claim that compulsive behavior, for instance, belongs to a different action-type by virtue of not being judgment-sensitive. Rather, it is the other way round: attributionist theories first have to give an independent account on how compulsive behavior differs from other kinds of intentional actions, and then explain how these differences support its supposed judgment-insensitivity.

Chapter 4: Attributionism and Carelessness

4.1 The Importance of What We Don't Care About

In Chapter 3 I argued that, contrary to Scanlon's claim, attributionist accounts cannot discriminate between the Frankfurtian addicts with respect to their responsibility. Now I would like to introduce another difference between hierarchical accounts (or mesh theories) and attributionism, which suggests that attributionist theories have crucial advantages over hierarchical accounts in defining the proper scope of responsibility-attribution.

Hierarchical compatibilist accounts, most notably Harry Frankfurt's and Gary Watson's theories, got much attention and came in for several serious strands of criticism in recent discussions about compatibilist strategies in the free will debate. Most prominently, the major arguments against them raised doubts about the success of explaining and modeling free and responsible agency by picking out a privileged part of human psychology (i.e., higher-order volitions or one's valuational system), with which the other parts should align. At the same time a different, although related problem, concerning the proper scope of responsibility-attribution, was rarely discussed. Watson himself admitted the difficulty, 12 years after having introduced his hierarchical proposal in "Free Agency" (1975). He writes,

> When it comes right down to it, I might fully 'embrace' a course of action, I do not judge best, but is fun, or thrilling; one loves doing it, and it's too bad it's not also the best thing to do, but one goes for it without compunction. Perhaps in such a case one must see this thrilling thing as good, must value it; but, again, one needn't see it as expressing or even conforming to a general standpoint one would be prepared to defend. (Watson 1987, p. 150)

The problem, at least for Watson's account, is that in these so-called "perverse" cases agents freely and responsibly refuse to act in accordance with their endorsed values, thus providing clear intuitive counterexamples to Watson's theory. But if we try to fix things by defining a person's valuational system as "what that person does, without regret, when it comes right down to it" (ibid.), we will end up, as Watson rightly recognizes, losing all the explanatory power of the concept of identification. Since both Frankfurt and Watson aim to provide not a formal model of intentional action, but a credible picture of human psychology complying with our pre-theoretical thinking, drastically redefining such concepts as "values" and "desires" is not a legitimate means for solving the problem of perverse cases.

I propose that the problem of perverse cases has much deeper roots, related to the concept of identification, and exploring these can raise doubts about hierarchical accounts in principle, not only about the versions presented by Frankfurt and Watson. The idea is that no matter to which part of human psychology we give privileged position and whether we define identification as conformity between the different parts or as control-exercising by the privileged part, there will be clear instances of responsible agency which will not show any sign of identification. Examples show a great variety and require case-by-case examination, but culpable ignorance and negligence, sudden emotional and behavioral responses as well as instances of practical irrationality, i.e., weakness of will[11] and self-deception certainly belong to those responsible actions and attitudes which cannot take their proper place within a hierarchical account.

The reasons why these familiar instances of human agency seem to be resistant to the hierarchical description of identification, are diverse. While in the case of culpable ignorance and negligence what lacking is conscious awareness in general, in the case of sudden emotional and behavioral responses we are often Frankfurtian wantons, by-passing any kind of higher order consideration; practical irrationality, finally, provides us examples of

[11] Watson has his own solution for akratic actions, but it would be beyond the scope of this discussion to present it.

disidentification due to the clash between judgments and other judgments or judgments and actions.

Nevertheless, despite their apparent diversity, the general moral that can be drawn from these examples is that with regard to responsibility-attribution, *not caring* about something is just as morally significant as caring about it. The picture which hierarchical accounts draw of human psychology is of such highly reflective human beings who express their concern about their own actions and attitudes through the process of constant self-evaluation and self-refinement. The question—which will determine the agent's responsibility—is only whether this process is instrumentally effective, i.e., capable of governing one's desires, attitudes and actions.

But this is often not the only relevant question. By way of an example, let's imagine someone who, despite that she knows that her partner has an ugly flew, does not spend more time with him. Although not being less kind or caring than usual, she does not ask about her partner's special needs, sees no reason for not having a fun night with her friends or to handle more patiently her partner's grouch.

Hierarchical theories have to face constant difficulties with explaining alike examples. It is clearly not the case that the agent attaches any value to ignoring her partner's sickness, and it would be even more implausible to say that she formed (either deliberately or spontaneously) a second-order desire to do so.

4.2 The Attributionist Solution

Attributionist accounts, by contrast, are designed to solve these kinds of cases. Intuitively, the reason why we hold the agent morally responsible and blameworthy is that she doesn't see her partner's sickness as a reason for caring more about him. And this is exactly what Scanlon's and Smith's accounts predict: the primary targets of moral criticism are the agent's judgments and lack of judgments (we will return to this distinction a bit later) as they are expressed in attitudes and actions. As Scanlon puts it: “A person can be criticized, and asked to provide justification or acknowledgment and apology, for things that seem to have been done inadvertently in a situation in which advertence is called for”. (Scanlon 1998, pp. 271–272.).

Smith also lays special emphasis on alike cases. In her paper “Responsibility for Attitudes” (2005) she discusses in length spontaneous attitudes and their central role in moral life. The titles of the subsections are telling in themselves: *Noticing and Neglecting*, *What Occurs to Us*, *Involuntary Reactions*. Smith gives an amazingly rich description of these everyday phenomena, which are usually undeservedly neglected in the philosophical literature of action theory and moral psychology. Smith demonstrates not only how naturally and obviously we respond to the moral qualities which are manifested in spontaneous attitudes—she also presents how the rational relations view, her attributionist account is able to explain moral responsibility for these reactions.

Let’s see then Smith’s key example: I have forgotten my friend’s birthday. Most would say that I am responsible and blameworthy for my forgetfulness. But how can we explain and justify this judgment of responsibility? Smith presents the problem in the following vain:

> But what, exactly, was the nature of my fault in this case? After all, I did not consciously choose to forget this special day or deliberately decide to ignore it. I did not intend to hurt my friend’s feelings or even foresee that my conduct would have this effect. I just forgot. It didn’t occur to me. I failed to notice. And yet, despite the apparent involuntariness of this failure, there was no doubt in either of our minds that I was, indeed, responsible for it. (Smith 2005, p. 236)

According to the rational relations view, the things which occur to us and those which we completely neglect; our general sensitivity or insensitivity to certain aspects of our environment can be proper subject of moral assessment because they are expressive of the agent's evaluative judgments about the weight and importance of these things. As Smith summarizes it:

> if one judges some thing or person to be important or significant in some way, this should (rationally) have an influence on one's tendency to notice factors which pertain to the existence, welfare, or flourishing of that thing or person. If this is so, then the fact that a person fails to take note of such factors in certain circumstances is at least some indication that she does not accept this evaluative judgment. (Smith 2005, p. 244)

Smith's argument goes as follow: if one holds a certain evaluative judgment that *x* is important, then she'll be disposed to notice relevant factors to *x*'s welfare. Thus, by using contraposition we can conclude that if someone is not disposed to notice relevant factors to *x*'s welfare, she does not hold the evaluative judgment that *x* is important.

So far, so good—the attributionist solution to the problem of involuntary omissions has some major advantages. First, as we have seen in Chapter 2, authors have constant difficulties with establishing the connection between responsibility and voluntary control in such cases. By simply denying the relevance of any such connection attributionist accounts relieve the discussion of this burden and offer a relatively easy solution for this puzzle. Second, just as in other cases, attributionist accounts do an excellent job in identifying the *content* of moral criticism. Indeed, it seems to be the case that I am blameworthy because I do not care enough about my friend, not because of any conscious or voluntary action or omission of mine. It is, just as in the previous example, the lack of concern which triggers moral criticism.

However, it was exactly the idea of such rational inferences (the very heart of the rational relations view) from the agent's attitudes or conduct to her evaluative judgments that came under fire recently. Matt King (2009) argues that such inferences are usually based on repeated evidences, while blameworthiness for behaving negligently does not presuppose any such regularities:

> The power of the evidential relation surely rests on the reliability of the inference from conduct to ill qualities of will. The reliability of such an inference requires, it seems, some regularity in its connections. (…) Of course, any conduct can count as some evidence for the underlying quality of will, but we generally require more before we are justified in actually drawing the inference. (…) But ascriptions of responsibility in cases of negligence need not rest on regularities. (…) [O]ne transgression is sufficient for negligence, and if negligence itself is to be sufficient for responsibility, then it seems that quality of will views (on the evidential reading) fare no better in explaining it, for the transgression itself won't be sufficient evidence for an ill quality of will. (King 2009, p. 584)

Holly Smith echoes King when she writes:

> In such cases, no *stable* faulty attitude could be attributed to the agent in light of his or her one-time failure to take notice. Indeed it may not even be plausible to ascribe to the agent a *momentary* faulty attitude of the kind shown by an exhausted soldier who shoots at a movement in the house he is searching, too tired to care about the risk that he is shooting an innocent civilian rather than an enemy combatant.
>
> In cases such as these, in which we can't reasonably impute a faulty evaluative attitude to the negligent agent, the Attributionist strategy for imputing blame to the agent for her culpably ignorant act seems to fail. (Smith 2011, p. 120)

4.3 Reasons Externalism and the Origination Thesis

In the last section I argued that while other theories (including hierarchical accounts) are rather perplexed by cases of negligence, carelessness, forgetfulness and the like, attributionism can provide a plausible answer: what makes the agent blameworthy is her judgment which is manifested in her spontaneous attitudes and behavior.

These notorious difficulties, which other accounts have to face, might be thought to be rooted on the adherence to the concept of control. While more traditional theories of freedom and responsibility focus on voluntariness and choice, mesh theories make use of some concept of identification or commitment—but both of them are equally incapable of handling cases where responsibility and blameworthiness are grounded in the *lack* of concern, care or attention.

But we need not be committed to any such notion to find these cases deeply problematic. In his subtle discussion of responsibility without awareness, i.e., cases where the agent is not aware of her wrongdoing, still apparently blameworthy for her behavior, George Sher (2009) recognizes a requirement, which we implicitly accept when searching for the conditions of moral responsibility:

> When we hold someone responsible for acting wrongly, the person whom we hold responsible is not the same as the feature of his act for which we hold him responsible. Because our blame and punishment are directed at an agent but are justified (if they are) by the wrongmaking features of what he has done, their grounding must include some appropriate relation between the agent and his act's wrong-making features. (Sher 2009, p. 147)

Sher argues convincingly that while the requirement of what he calls the *origination relation* "is forced upon us by a deep structural fact about responsibility" (ibid.), the idea of voluntary control as a prerequisite of responsibility is only one possible interpretation of this more abstract truth. Consequently, one can legitimately refute the Control Principle, but in one way or another has to establish an appropriate origination relation between the agent and her action's wrong-making features.

I suspect that it is the lack of this crucial connection that most authors find troubling in cases of involuntary omissions. Since in all cases we are talking about *omissions* coming about without the agents being *aware* of them, it is mysterious how they can be said to originate in the agents. Also, this is a requirement which attributionist accounts cannot bypass by refuting the role of voluntariness and control in establishing judgments of responsibility—Sher's point is exactly that the origination relation cannot simply be ignored even if we reject the Control Principle. It would exceed the limits of this discussion to present Sher's own proposal. Instead, let's explore how attributionist accounts face this challenge.

This is one of the rare occasions, when it is important to discuss Scanlon's and Smith's accounts separately. While Smith talks about judgments being sensitive to one's evaluative judgments in a loosely defined sense, Scanlon discusses judgments about reasons, and this "reason-talk" brings up some well-known philosophical issues.

For the first sight the problem of origination relations seems to be closely related to the debate about the existence of external reasons. The first formulation of the question was put forward by Bernard Williams (1979). In his now classical example Owen Wingrave is pressured by his family to follow the family tradition and pursue a military carrier. Owen, however, "hates everything about military life and what it means" (Williams 1979, p. 106). The question is straightforward: are Owen's relatives right to think that Owen still has a reason to join the army?

Williams answer is no.[12] Reasons in the normative sense have to be able to *explain* the agent's behavior: the agent's action has to intelligible in the light of his subjective motivational set. Since in this case there is no "sound deliberative route" leading from elements of Owen's subjective motivational set to the conclusion that there is something counting in favor of joining the army, we cannot say that he has any reason to join the army. To put it in more

[12] At least this is how most interpreters understand him.

general terms, reasons internalists (including Williams) claim that “a consideration is a reason for an agent only if some motivational fact about that agent obtains” (Finlay 2008). Reasons externalists, by contrast, maintain that there can be other kinds of reasons as well beyond those internalists allow for.

The disagreement between reasons internalists and externalists has direct relevance when it comes to moral and non-moral judgments. In Owen’s case his relatives will most probably complain if he does not join the army and will predictably judge his decision rather unfavorably. But, if reasons internalists are right, then these criticisms cannot be based on the charge that he did not recognize his reasons for joining the army—because he did not have any reason whatsoever for so acting. Or, closer to home, let’s return to our previous example of the woman who does not take care of her partner. Neither internalists, nor externalists deny that our agent can have a reason for paying special attention to her partner. But, according to the internalist, in order to have a reason for that the agent has to have something in her subjective motivational set, e.g. a general concern for other people’s well-being or a desire to see her partner healthy again, which can lead to the conclusion that there is something to be said in favor of paying special attention to her partner. Externalists will say, by contrast, that one can have this reason also in the absence of such elements of her psychology. And the burden of proof is on them to show what makes such reason-attributions appropriate.

According to Scanlon’s analysis our agent is blameworthy because she does not judge her partner’s sickness to be a reason counting in favor of paying special attention to him, or, alternatively, because she judges that her partner’s sickness is not a reason for paying special attention. However, once we take an internalist position it becomes mysterious what is wrong with this—since she does not, in fact, have such a reason. Why should we blame someone for not taking something as a reason if for her it is not a reason?

Scanlon clearly recognizes the problem, when he writes about a like example: “If it is a deficiency for the man to fail to see these considerations as reasons, it would seem that they must be reasons for him. (If they are not, how can it be a deficiency for him to fail to recognize them?)” (Scanlon 1998, Appendix, p. 367) To defend his account, he is thus pushed toward the externalist position which he seems to accept with minor qualifications. According to his theory moral reasons are determined by those principles which nobody similarly motivated would reject. Not recognizing moral reasons can be adequate basis of moral assessment because they are the agent’s reasons no matter what elements her subjective motivational set contains.

Reasons externalism, however, is an untenable view if one is to explain responsibility in terms of judgments about reasons and wishes to account for the origination relation. If a strong enough connection between the agent and her action’s wrong-making feature is necessary to establish judgments of responsibility, then we cannot explain the act’s wrongness by referring to reasons which are completely independent of the agent’s subjective motivational set. No matter how we conceive external reasons, it is clear that they are incapable of establishing the origination relation.

Does Smith’s rational relation view succeed any better in answering Sher’s proposal? Independent examination is justified since the notion of reason does not play any role in Smith’s account. Instead, Smith talks about evaluative judgments, which gives a more subjective twist to attributionism. The central difference is that while Scanlon uses the term “reason” in a normative sense, for Smith having an evaluative judgment is a descriptive psychological fact. That is, although evaluative judgments due to their belief-like structure can be unconscious, they are real, existing parts of the agent’s psychology, contrary to reasons, which can exist by virtue of some possible “sound deliberative route”, which the agent has never taken or, as in the case of Scanlon, by virtue of the truth of some propositions about right and wrong.

In the previous subsection we explored Smith's solution for the problem of involuntary omissions. When explaining the forgotten birthday example Smith argued that since a judgment about *x*'s importance is, in an ideally rational world, expressed, among other things, in the agent's noticing important factors to *x*'s welfare, from the lack of noticing we can reasonably infer that she does not hold an evaluative judgment about *x*'s importance. This is why we blame her—in Sher's term, this is the wrong-making feature of her omission.

Is this an adequate solution of the puzzle of responsibility for involuntary omissions? Again, what is especially troubling in these cases is that we cannot establish any connection between the agent and her omission, which would make holding her responsible justified. However, I argue, Smith's argument makes no progress in this respect—it merely shifts the problem to another region. Instead of talking about failing to notice something we should have noticed the problem is now that we do not hold a judgment which we should hold. In order to tie the agent to her failure what we should show is not that the agent does not hold a judgment about *x*'s importance, but that she does hold a judgment of *x*'s unimportance. Without that it is rather mysterious what the target of responsibility-attribution is, since we couldn't make any positive claim about the agent. But this claim simply does not follow from Smith's argument.

This is structurally the same problem as the one which Sher recognizes, when he talks about an account which closely resembles Scanlon's view:

> What we are trying to understand is how an agent can be responsible for performing a wrong or foolish act that he did not recognize *as* wrong or foolish. The explanation we are considering is that what makes such an agent responsible is his failure to recognize the force of his reasons for believing that the act *is* wrong or foolish. But if there is a problem about how an agent can be responsible for

> acting wrongly or foolishly despite his failure to recognize that his *act* is wrong or foolish, then isn't there just as much of a problem about how an agent can be responsible for acting wrongly or foolishly despite his failure to recognize his *reasons for believing* that his act is wrong or foolish? If we are trying to link the act's wrongness or foolishness to the agent in some suitably strong sense, then how does it help to relocate the point at which the crucial failure of recognition occurs? Even if we replace the claim that the agent has failed to recognize his act's wrongness or foolishness with the claim that what he has failed to recognize are the reasons that his situation or other beliefs gave him for believing that his act *was* wrong or foolish, won't it remain true that the act's wrongness or foolishness played no role in his practical reasoning? (Sher 2009, p. 79)

Now we can see that it is not primarily Scanlon's commitment to external reasons that generates problems for attributionist accounts. The problem is more general: if we take Sher's requirement of the origination relation seriously (and I suspect that refusing it would require fundamental changes in our thinking about moral responsibility), then we have to be able to identify a positive claim about the agent's psychology which explains why she failed to notice, remember, take care, etc., what she should have noticed, remembered, taken care of, etc. Attributionist accounts, despite their apparent terminological complexity seem to merely reiterate the claim that something that should have been done has not been done.

At this point Scanlon would oppose. In an interesting, although somewhat ambiguous passage he writes: "A person who is unable to see why the fact that his action would injure me should count against it still holds that this doesn't count against it" (Scanlon 1998, p. 288). Unfortunately there is no wider context which would explain or specify this claim. What seems to be more or less clear is that in a generalized form the claim is that not judging something to be a reason is equivalent to judging that it is not a reason. If this thesis were true, it could save attributionist accounts from failing to establish the origination relation between the agent and her failing's wrong-making feature. In the case of the forgotten birthday, for instance, Smith could say that the agent’s failure to notice a relevant factor to her friend’s wellbeing manifests her judgment that her friend is not important.

Since Scanlon does not provide any argument supporting this claim, it is our task to test its feasibility. As a starting point I would return to one of my previous points, i.e., that despite their peculiar terminology, judgments in both Scanlon’s and Smith’s theories have a belief-like structure. Thus the question is whether it is reasonable to infer from the agent’s attitudes and actions not only to the lack of a belief, but to the possession of a belief with contradictory content.

In some cases it can be. To take an example: I have never in my life played Scrabble on the middle of a highway. I don’t have a desire to try it, I have never intended to do it and I don’t feel happily excited when thinking about it. Based on these attitudes of mine we can reasonably say that I don’t have the belief that “Playing Scrabble on the middle of a highway is good/enjoyable/etc.”, since if I had, I would be disposed in many ways to manifest it. But can we say that I do have a belief that “Playing Scrabble on the middle of a highway is not good/enjoyable/etc.”?

In some sense we can. If, for instance, someone asked me whether I think that playing Scrabble on the middle of a highway was good or enjoyable, I would protest without hesitation. However, there is still something unpleasantly artificial to say that we hold all these beliefs about subjects of which we do not even think. It sounds odd mainly because one of the most palpable manifestations of beliefs is judgment-making, i.e., the conscious assent to them.

Thus, although it is strictly speaking not false that I hold a belief about the badness of Scrabble-playing on the highway, this claim sounds utterly counterintuitive. This becomes even more apparent when we consider other cases, which are structurally indistinguishable from the previous example, yet clearly do not allow for making such inferences. For instance, for long decades the though of playing curling has not once came into my mind. I didn't have a desire to play curling, never intended to do so and did not feel happily excited when thinking about it. But although it would have been reasonable to infer from my attitudes and conduct that I do not believe that curling is fun/enjoyable/etc., it certainly wasn't the case that I thought curling not to be fun/enjoyable/etc. I simply did not have an opinion on that matter, that's all.

In sum, I find Scanlon's claim in its generalized version highly contentious, to say the least. Thus, we are back to where we started. It seems that sometimes we do hold people responsible on the basis of not recognizing some of their (supposedly moral) reasons for action. The question is how to describe their behavior by referring to such positive psychological facts which would establish an appropriate relation between the agents and their failure to recognize certain reasons. In this subsection I was arguing that attributionist accounts don't make such a description available.

Chapter 5: Responsibility for Emotions – Angela Smith's Account

Our intuitions supporting the view that we are held responsible for our emotions, are nicely summarized by Eugen Schlossberger who analyzes the case of Charlie who enjoys watching the suffering of animals:

> it seems impossible to deny that it would be a moral improvement were Charlie to stop enjoying the sight of animals in pain. Again, surely someone who is otherwise just like Charlie but does not enjoy the torments of animals is morally preferable to Charlie. So Charlie would be a morally better person were he to stop taking such sadistic pleasure. (Schlossberger 1985, p. 40)

In a similar vein, it is pretty common to blame someone for envying one's friend's success or feeling aversion toward certain minority groups. Note that this problem is distinct from issues of responsibility for acting on an emotion or expressing an emotion—the topic I will discuss is whether merely having (or, less neutrally, feeling) an emotion can be the basis of moral evaluation and if it can, on what grounds.

Let's see the passage where Smith characterizes the relationship between evaluative judgments and emotions:

> Attitudes such as contempt, jealousy, and regret seem to be partially constituted by certain kinds of evaluative judgments or appraisals. To feel contempt toward some person, for example, involves the judgment that she has some feature or has behaved in some way which makes her unworthy of one's respect, and to feel regret involves the judgment that something of value has been lost. There seems to be a conceptual connection between having these attitudes and making, or being disposed to make, certain kinds of judgments. This helps to explain why we attach so much significance to these reactions, both in our own case and in our relations with others: unlike brute sensations, which simply assail us, our spontaneous reactions reveal, in a direct and sometimes distressing way, the underlying evaluative commitments shaping our responses to the situations in which we find ourselves. (Smith 2005, p. 250)

As we can see, Smith joins a long philosophical tradition of classifying emotions on the basis of their constitutive beliefs or judgments. Judgmentalism claims that some kind of cognitive component is necessary, although according to the more plausible accounts not sufficient, in order to experience an emotional state. Thus, Smith's argument continues, on the basis of observing an emotion on the agent's part we can reasonably infer that she holds a certain evaluative judgment.

I would like to raise two difficulties regarding Smith's account of responsibility for emotions, only one of which I will elaborate on. First, it is far from clear that Smith's concept of judgment, i.e., a "continuing and relatively stable dispositions to respond in particular ways to particular situations" (Smith 2005, p. 251, fn. 27), is an adequate candidate for the role of the cognitive component of emotions. It can be doubted whether a stable disposition is always constitutive to an emotion. That is, when we are faced with a completely new experience (tasting a new food or being challenged by a previously unknown moral dilemma), our emotions reflect such evaluative judgments which couldn't have been previously established. The only way to resist the conclusion that the cognitive components of emotions are rather temporary judgments than long-standing dispositions would be to argue that when we are faced with new experiences and new emotional responses, what constitutes our emotion is a long-standing disposition of a general kind, which, by inference, is applied to the new cases.

But even if we set this problem aside, there is another one concerning the scope of Smith's account. Contrary to the rational relations view, it is not true that since emotions are in principle sensitive to evaluative judgments (whatever they may be), we are responsible for all our emotions. I think that our moral intuitions are pretty strong in telling that there are at least two exemptions which Smith seems to ignore, although she admits the existence of them.

The first are recalcitrant emotions, more precisely phobias. Having a phobia provides an excuse for my irrational fear, and in most of the cases claiming that one has a phobia amounts to denying one's responsibility for her fear. Something similar happens in the case of the sudden coming-into-existence of basic emotions such as fear or surprise. No one would assess me (either morally or non-morally) for getting frightened of a loud, unexpected noise or sight. Sudden and basic emotional responses fall into the same class as involuntary bodily movements (which may be because basic emotions are often in fact accompanied by involuntary bodily movements)—we simply do not raise the question of moral evaluation on the basis of them.

However, despite its apparent inadequacy, attributionism seems to get things basically right. Both in the cases of phobias and sudden basic emotions, the intuitive reason why we do not hold people responsible for them is exactly their assumed judgment-insensitivity. So one obvious strategy for saving the attributionist account would be to reject Smith's type-attributionism and apply the condition of judgment-sensitivity once again for attitude-tokens. But with this move we would simply beg the question against the judgmentalist. Token-attributionism, according to my suggestion, would be then committed to roughly the following condition:

> One is morally responsible for an emotion if it is true that if the agent had held a different set of judgments, the emotion wouldn't have come into existence.

But if we accept judgmentalism, then this will be true of every emotion. Since being a judgmentalist means exactly to exclude the possibility of an emotion's coming-into-existence without the appropriate cognitive component, within a unified judgmentalist framework we are not able to discriminate between phobias and sudden basic emotions on the one hand and other emotions, on the other hand, on the basis of their sensitivity to judgments.

I find that this tension, i.e., between the apparent plausibility of attributionism and its failure to make the supposed distinction between judgment-sensitive and judgment-insensitive emotions, suggests that the unified judgmentalist theory, which Smith accepts, has to be given up. In the following paragraphs I will briefly introduce an alternative account of emotions introduced by D'Arms and Jacobson, which can do the work judgmentalism can't.

D'Arms and Jacobson distinguish two kinds of emotions: natural emotion kinds and their cognitive sharpenings. Natural emotion kinds, which include emotions traditionally classified as basic and supposedly some more, are "products of relatively discrete special-purpose mechanisms that are sensitive to some important aspect of human life" (D'Arms and Jacobson 2001, p. 138). One of the main characteristics of natural emotion kinds is that they can function relatively independently of the agent's linguistic mechanisms.

Cognitive sharpenings, by contrast, are "constructed by specifying a subclass of instances of an emotion, or other affective state, in terms of some thought that they happen to share" (Ibid, p. 137). Examples include homesickness (sadness invoked by the thought that one is far from home), resentment (anger over someone else's moral wrongdoing) or so-called tenure-rage, the authors' invention, which is the anger one feels when she was denied tenure. As this last example shows, there can be infinitely many cognitive sharpenings depending on the thoughts by which we classify them. D'Arms and Jacobson argue that while judgmentalism is right with respect to cognitive sharpenings, it is simply false with regard to natural emotion kinds. Although natural emotion kinds doubtlessly serve certain purposes and thus have a strong connection to such concepts as danger (in the case of fear) or contamination (in the case of disgust), it does not follow from this that "in order to feel fear, one must deploy this or any other concept" (ibid., p. 139). On the contrary, natural emotion kinds often come into existence by by-passing the agent's linguistic mechanisms.

Hopefully this very schematic sketch of D'Arms and Jacobson's account of emotions can shed light on how the attributionist explanation of responsibility for emotions can be saved. By implementing D'Arms and Jacobson's idea we can maintain that while natural emotion kinds such as fear (recalcitrant or not) and surprise are produced by mechanisms independent of one's linguistic system and consequently are not sensitive to one's judgments, cognitive sharpenings are sensitive to their constitutive thoughts and thus can be the basis of moral assessment. Judgmentalism failed because it regarded emotions as members of one homogeneous class to which in principle judgment-sensitivity applies to. If we accept the distinction between natural emotion kinds and cognitive sharpenings, we will be able to discriminate between two types of emotions, only one of which in principle judgment-sensitivity applies to.

Note that D'Arms and Jacobson's list of natural emotional kinds include much more emotions than what I have mentioned. According to them envy, jealousy and shame also belong to natural emotional kinds, although we are commonly assessed on the basis of them. However, as the authors also admit it, this list is only provisionary and it is a matter of scientific investigation to find out which emotions enjoy typically the kind of independence from the linguistic mechanism which D'Arms and Jacobson talk about. Depending on the empirical results, our theory of responsibility can be very conservative as well as highly revisionary. If it turns out that guilt and jealousy are usually not sensitive to our judgments, then up to now we followed an inappropriate practice when we held people responsible on the basis of them and thus should give this practice up. By contrast, if we discover that our emotional experience of fear is tightly connected to one's linguistic system, then in the future we can be harsher in assessing people for it. I take this indeterminacy of the theory as an advantage of it, since it can explain our fairly common reluctance or hesitance to attribute to the agent e.g. her excessive jealousy or permanent disgust of something.

Chapter 6: What Do We Mean By Control?

In this chapter I undertake a task which is largely independent from attributionist theories, but closely related to the clash between the Control Principle and judgments of responsibility in problematic cases. I will present yet another strategy to dissolve this clash, that I find particularly promising. I would like to show that, given how we ordinarily use the notion of control, we have positive reasons to suppose that the Control Principle does not contradict to all problematic cases. Note that I do not intend to give a uniform treatment to all cases. As a matter of fact I am convinced that whatever the most suitable interpretation of the Control Principle is, it *is* violated if resultant and circumstantial luck cases really do obtain. Moreover, I have no solution whatsoever for these cases. However, since the focus of my investigation is the feasibility of attributionist accounts, I can legitimately restrict my attention to those cases which *they* aim to explain: responsibility for attitudes and negligent behavior. Also, I will not provide an alternative concept of control: I will only explore what we usually *mean* by control and in what contexts we use it (or its lack thereof) to underline or deny one's responsibility for something.

6.1 Losing Control, Beyond Control, Out of Control

First, we should recognize that most often we talk about exercising control over things *external* to us. Most evidently, we control tools, vehicles and machines (it is no surprise that driving a car is a typical example in the literature of control). We control them because it depends on us how they function; their functioning reflects our wants, needs and intentions. Interestingly, the degree of control which we have over these things does not depend on the complexity of their functioning. That is, it does not matter whether my hair dryer has six or only one speed setting—my control over it is just the same (the former is a better hair dryer, though). Also, our ability to control something does not depend on whether we actually control it. I have just as much control over which channel is running on my television when the remote control is in the other room as when I am pushing its buttons. The point I'm pushing here might sound somewhat trivial: I do not lose my control over my recorder every time when I do not use it.

Still, it is sensible to say when, for instance, sitting on a talk at the department building, that unfortunately I cannot control my video recorder from this distance. By saying this I can express three different things: (i) a disability of mine: although I have downloaded an iPhone application which serves exactly this purpose, I still don't know how to use it, or my hands are paralyzed so I cannot use the touch screen; (ii) some kind of malfunctioning on the part of the machine: the application is supposed to be working, but for some reason it cannot communicate with the recorder (this happens embarrassingly often, by the way); or (iii) in principle impossibility: recorders simply cannot be programmed from a larger distance. These three ways of not having control are significantly different and function differently in excusing or exempting someone from responsibility.

Let's suppose that it would be important (for some unknown moral reason) to record a TV show and I am the only one available for the task. However, as we know, I am sitting at a talk and there is no opportunity for me to leave. How do the aforementioned three types of lack of control influence my responsibility, if finally I don't record the program?

The last case seems to be the most obvious: it seems unreasonable even to address the demand of recording the program, given that I am at the department and recorders cannot be programmed from this distance. Just as I cannot fly or lift 400 pounds, I cannot do anything to advance the desired aim. In this sense it is adequate to say that the thing in question is beyond my control.

Things are different in the two other cases. In the second case it can rightly be said that I have *lost* control over my recorder, although previously I did control it. Although this is most probably a successful excuse, further questions can be raised: did I previously notice that the device is malfunctioning? If yes, did I have any reason to think that the malfunctioning would prevent me of fulfilling a moral demand? If not, should I have known? These further questions indicate that although losing control does serve in many cases as an excusing condition, contrary to the first case, here the demand itself is not unreasonable at all.

The same goes for the first type of cases. Although my inability to control the application makes it impossible for me to perform the desired action, it can rightly be ask whether this inability is explained by some more basic incapacity of mine or is it the result of simple carelessness. But, if the former is true, that might have similar consequences as in the first type of cases: if I cannot use the application, because due to some deficiency I am generally incapable to learn how to use such devices, then addressing a demand of this kind *to me* is unreasonable (although it might be perfectly reasonable in case of people with normal capabilities).

6.2 Attitudes

How do these observations help us to understand problematic cases of responsibility? Most importantly, it is crucial to define in exactly which sense we lack control over our attitudes and negligent behavior. First, let’s examine attitudes! It seems to me that both the first and the second form of denying control are unpromising candidates, if we want to make sense of the claim that it is unfair to hold agents responsible for these instances of agency. Granted that we don’t want to say that the lack of control over attitudes is to be explained either by a disability of certain agents (which wouldn’t make responsibility-attribution universally unfair) or by referring to some kind of malfunctioning (but how could *all* attitudes be anomalies?), then the only solution left is to say that attitudes are not the kind of things which we could, even in principle, control. Note that, just as in the case of the recorder, this claim calls into question not only the fairness of responsibility-attributing practices, but also any kind of moral demand which would require us to have an attitude, but not another.

However, claiming that we are entirely unable to control our attitudes seems to be mistaken. Here I cannot go through all the attitudes—emotions, beliefs, intentions, desires, character traits—for which we frequently hold people morally responsible. My remarks will be thus restricted, in an entirely *ad hoc* way, to emotions and beliefs. I contend that my main point applies with minor modifications also to other types of attitudes, but here I won't even make an attempt to argue for this claim.

Let's start with emotions. It seems that there are several types of methods by which we can tame, change, mitigate and even eliminate our emotions. First, there are methods which might be labeled as therapeutic. We can temper our anger by working out in the gym; reduce our feeling of disappointment by focusing on new, hopeful challenges; chase away our sadness by meeting our friends or watching romantic comedies. These therapies do not affect the causes and reasons for having the emotion, but divert our attention and energy so that the emotional experience becomes less intense.

Second, we can change or eliminate our emotions by changing our beliefs, since many emotions depend importantly on how we see the world. We try to eliminate our jealousy by finding out whether our partner is loyal to us (although there is a chance that the result will not help in removing our feeling); or mitigate our resentment by considering the other's intentions and available alternatives by the time of her faulty action. This second method might be effective in removing the emotion altogether: if it turns out that our initial diagnosis of the states of affair was mistaken, then the emotion will most probably vanish.

Finally we can, in an attributionist fashion, modify our emotions by changing the evaluative commitments constitutive to them. The difference between this method and the previous one can be illustrated by the following example. Let's say I'm afraid to go to a party because I suspect that my ex-boyfriend will be there. While the previous method requires me to find out whether he really would be there, and eliminate my fear if it turns out that he wouldn't, this last method requires me to realize: there is nothing the least dangerous in meeting my ex. In a similar vein, I can reduce my contempt toward other people by respecting that there are other, equally valuable life styles than my own; or my feeling of being betrayed by my friend on reflecting what friendship requires. Surely, these methods are not *always* effective: as we have seen in the discussion of Angela Smith's account, emotions are often stubborn and recalcitrant. But it would also be an exaggeration to say that the results are completely chancy—by using these indirect methods we can reliably get rid of (or at least temper) unwanted emotions.

Controlling beliefs, at first glance, seem to be even more complicated, since most often we acquire and hold beliefs in an unconscious and automatic level. Given the perceptual information provided by my environment I cannot help thinking that right now there is a brown desk in front of me and that it is raining outside. Moreover, as many argue (on the aims of beliefs see e.g. Owens 2003, Velleman 2000 and Whiting 2012), beliefs have a specific, inherent aim: tracking truth, and this characteristic seem to prevent us to believe just whatever we want to believe. Still, we possess several means to form and revise beliefs. On the one hand, we might use methods similar to what I called "therapeutic" in the case of emotions: we can engage in self-deception or, to the extreme, use hypnosis or narcotics to acquire the beliefs we want. However, these methods can succeed without even considering the truth or falsity of the given belief and for this reason they might be considered as rather unusual ways of acquiring beliefs. But there are several other methods which respect the truth-tracking nature of beliefs: these can be actions—collecting evidences—but also involve several mental activities such as assessing and balancing evidence, deliberating, guessing or coming to a conclusion. Again, I do not want to claim that these are *perfectly* reliable methods: we sometimes feel being stuck with a belief, which is not supported by sufficient reasons (or by any reason at all). Recognizing these cases, some authors (see e.g. Owens 2000) concluded that we do not have any kind of rational control over our beliefs. I resist this conclusion: beliefs lacking any good reason construe a similar kind of malfunctioning as recalcitrant emotions and should be handled as exceptions rather than rules.

Now that we have seen the varieties of methods effective in modifying and eliminating emotions and beliefs one might ask why anyone would deny that we can control them. I find that the resistance to admit that we can control our attitudes arises from two different sources. The first, which I tried to highlight in the previous paragraphs, is the supposed unreliability of the aforementioned methods. The common experience of being passive and helpless with regard to our beliefs and emotions led many philosophers to deny any possibility to exercise control over them, or, more moderately, to claim that our control and accordingly our responsibility for them can only be partial (see Ben Ze'ev).

Although real disputes about the reliability of the methods I have presented are at place here, and it might be difficult to do justice between those who are optimistic about the prospects of these methods and those who are rather pessimistic, I would like to emphasize one point which is independent of these disputes. The mere fact that we have non-chancy methods for modifying, mitigating and eliminating beliefs and emotions provides sufficient counterargument against the original claim. Remember, those who insist that there is a genuine tension between the Control Principle and the moral assessments of attitudes, want to establish that beliefs and emotions are in principle beyond our control. We do not have to present a perfectly or almost perfectly reliable method of controlling our attitudes: simply having such methods makes this claim false. Surely, that these methods are not perfectly reliable shows that time after time we *lose* control over our attitudes: but such instances of malfunctioning do not question the general ability of control.

Second, more importantly, doubts about the possibility of controlling attitudes often come from a comparison between attitudes and actions. That is, no matter how much I try, I won't be able to get rid of any of my beliefs or emotions the way I am able to raise my arm or cross my legs. We cannot induce, modify or eliminate attitudes at our will: contrary to actions, we cannot simply choose what we want to do with them and then do that instantly. If the paradigm case of control-exercising is controlling our actions, then the influence over our attitudes certainly fall short of what might be called control.

It might be unnecessary to emphasize how central these ideas are in our thinking about freedom and responsibility. Neither is it requisite to list all those theories according to which without the notions of will and choice we cannot tell anything useful about why human action is different and distinguished from mere happenings and the actions of non-humans. Here instead I would like to return to one of the initial points which I raised in Chapter 2. The concept of control plays a prominent role in theories of responsibility and many authors assume that it is a necessary condition of responsibility. For this very reason it is a crucial task to give a viable analysis of the concept. However, if we take it seriously that the notion of control is to explain why responsibility-attributions are appropriate in certain cases, but not in others, then the concept has to be developed independently of intuitions about responsibility. Otherwise the notion of control will have no explanatory power, and at the end of the day the

Control Principle will boil down to the moderately interesting truth that it is unfair to hold someone responsible, if in fact she is not responsible.

It might be that actions preceded by conscious deliberation and choice are the paradigm examples of free and responsible agency. But this gives us no reason whatsoever to think that they are also the paradigm cases of control. My previous examples and observations tried to motivate a shift of attention from control over our actions to control over external objects. Obviously, I do not want to deny that we *do* control our actions. Moreover, our control over our bodily movements is arguably more nuanced and perfect than any other kind of control which we can possess. However, the moral of exploring other usual contexts in which we talk about control is that in order to have *full* control over something we do not have to have *perfect* control. The degree of my control depends on my relevant abilities and the proper functioning of the machinery which I want to control. As long as I have all the abilities necessary to understand how to control the machinery, necessary for the execution of this control and the machinery functions properly, my control over the machinery is *full*, no matter how primitive it is.

It is a wholly different question how *perfect* my control is. In general, we are interested in developing our environment so that the things we control would become more sensitive to our wants and needs. We like it that we can set the exact heat of our oven; that our clock alarms exactly at the minute we want it to alarm. Usually, the more complex the machineries are, the more perfect our control is over them. Also, it is true that by making these objects more sensitive to our wants, we gain control over more things. To return to my original example: if my hairdryer has only one single heat setting, then I cannot control the temperature of the air it blows—it is beyond my control to do so. But this does not mean that I cannot control the hairdryer itself, or that I would have a fuller control over it if it had more heat settings.

The distinction introduced in the previous paragraphs can help us to illuminate the problem of controlling attitudes. If we start considering examples of control-exercising other than actions, it becomes clear that we have set the bar too high. We can rightly be said to control something even if our intention cannot directly and right away lead to the desired aim. I control the speed of my car, even though I am not able to stop it in an instance when driving with 100 km/h. To a limited, but quite significant extent we control our weight, although our decision to lose weight will lead to result only after a significant amount of time and effort. Doing directly and right away what one has decided to do is a peculiar and amazing characteristic of acting—but it is not a necessary component of control.

Earlier in Chapter 2 we encountered indirect or tracing theories. Now one could say that by describing the control we exercise over external objects I simply presented yet another—in its present form rather crude—version of tracing theories. After all, all the varieties of control I have mentioned so far are derivative from our control over actions. We control machines, vehicles and also beliefs and emotions by performing actions which we choose. How on earth would we exercise control if not by taking steps, pushing buttons, saying this instead of that? All the direct control that we have is over our basic actions, so, it can be argued, talking about control over anything else is only an inaccurate and misleading way of talking, which we shouldn’t seriously consider when exploring the nature of control. If we are to explain responsibility for attitudes we have to go back to those actions, by the help of which we maintain, modify or eliminate them.

I cannot give a conclusive answer to this objection here, but some important remarks are at place. First, as I said at the beginning of the discussion, it is beyond my present aims to offer an alternative concept of control. My goal is to analyze those contexts, in which we ordinarily attribute control to someone and broadly define our reasons for doing so. Since the Control

Principle gains it appeal from such an ordinary, non-interpreted concept of control, I find the task of conceptual analysis legitimate and useful. It might be that after rigorous philosophical investigation we will come to the conclusion that the only (metaphysically or morally) relevant form of control is control over our actions. In everyday communication, however, it makes perfect sense to talk about controlling external things. And we have no reason to think that the Control Principle relies on another, philosophically more elaborated or sustainable notion.

Second, while describing the familiar methods by which we maintain, revise and eliminate our emotions and beliefs, I made several references to *mental actions*, as opposed to "ordinary" ones: balancing evidence, deliberating, guessing, assessing, coming to a conclusion, judging. Describing such mental actions is a fascinating philosophical task which got significant attention only recently (see e.g. O'Brien & Soteriou 2009). However, even these first, cautious attempts to characterize certain mental actions suggest that they significantly diverge from ordinary actions with respect to our control over them (see e.g. Hieronymi 2006, 2009). Also, it is extremely implausible to suppose that our control over these mental actions can be traced back to ordinary actions.

And third, this proposal has serious deficits in characterizing and explaining control in general. Even if we accept that control over our actions is a necessary condition for controlling external objects, it is evidently not a sufficient one. As my examples made it clear, the ability of control does not depend solely on the abilities of the agent—it is also the internal structure of the given thing which makes control exercising over it possible for us. We cannot explain why we exercise control over cars by referring only to our ability to step on the gas or change the gears. We are able to control them because cars are designed in a way which makes them sensitive in a specific way to our bodily movements. Similarly, the reason why a boss can control her employees is not that she can talk, write and walk from one place to another (although she might lose her control without these abilities): her control depends on the hierarchical structure of the corporation which enables and entitles her to give orders, collect information and so on. Since claims about the incompatibility of the Control Principle and certain responsibility-attributing practices often hang on whether certain instances of human agency, such as beliefs and emotions, are in principle beyond our control, by reducing every kind of control to action control we won't be able to assess these claims—we simply beg the question against those who argue that we can exercise control over our attitudes.

6.3 Negligence

Cases of negligence raise different concerns about control than attitudes. Authors who deny that we are morally responsible for instances of absentmindedness, carelessness, etc. do not want to claim that noticing something, keeping something in mind, paying attention to something are things for which we can never be held responsible. Interestingly, the question whether these activities in general are within our control got little theoretical attention. It seems that authors assume some kind of basic asymmetry between the execution of these activities and the failure to exercise them. I find this assumption implausible, and, in the absence of further arguments, certainly premature. As we will see, the problem of negligence is tightly connected to the idea of *conscious* control. However, I see no reason to suppose that we exercise (or have exercised) more conscious control when we succeed in keeping something in mind, paying attention, etc. than when we fail to succeed in doing so. The characterization of these semi-automatic, semi-voluntary responses pose a major challenge to any theory of responsibility which gives role to conscious reasoning and choice in determining the agent's responsibility.

Previously I distinguished three kinds of reasons which make it appropriate to say that someone doesn’t have (or lost) control over something: (i) the agent’s inability to control the given thing (due to either ignorance or incapacity to perform the required action), (ii) the malfunctioning of the given thing and (iii) the thing’s internal structure doesn’t make control exercising in principle possible. Since we rejected (iii), we have two more options: negligence is caused either by some kind of inability on the agent’s part or by some kind of malfunctioning.

Let’s consider the latter option: can we say that the lack of attention or due care is some kind of a malfunction? To answer this question first we have to know the malfunction of *what* we are discussing. Analogies with external objects are of no help anymore, since, if anything, negligence is a malfunctioning of *us*. And, morally speaking, this is right: not knowing, recognizing or remembering what one should know, recognize or remember is a malfunction in the sense that in some way or another we fail to meet a moral expectation or obligation. However, this is obviously not the sense of malfunctioning which we should take into account, unless we want to say that every kind of moral failure is due to some malfunctioning on our part. This would be equivalent to saying that we are *never* in control (and so not responsible) when our behavior do not conform to the moral expectations—a radical conclusion few would accept.

But the “should” in question might also mean something weaker: given the agent’s aims, desires and beliefs, the negligence *normally* would not occur, so it has to explained by some kind of malfunctioning. Arguably, if one cares about someone or something, then she will be disposed to notice and keep in mind relevant facts about the given person or thing. Negligent behavior arises, an alternative proposal would say, when this rational connection is blocked due to external reasons.

Obviously, this idea goes straight against the attributionist view, which maintains that negligent behavior usually expresses one's evaluative commitments—this is exactly why we can rightly attribute it to the agent. Also, we have no reason to accept the claim that negligence always involves a discrepancy between one's evaluation and conduct. Probably it is worthwhile to briefly return to our previous example of Alessandra, who forgets that her dog is waiting in the hot car. There is no hint in the story which would suggest that Alessandra's failure to notice that Sheba is in danger would be due to any kind of disregard or general carelessness toward the dog. But we can just as well imagine an alternative scenario, where Alessandra's failure to take notice of Sheba is in perfect harmony with her views about the relative unimportance of animal life or her low opinion of Sheba particularly. Although in this case there would be no inconsistency involved, Alessandra's behavior would still remain an instance of negligent behavior. Negligent behavior, *pace* Scanlon and Smith, might come about with or without evaluative commitments being manifested in it—thus we cannot discriminate on this basis.

More importantly, it is unclear why Alessandra would be out of control in the first place. Surely, her attention has been diverted from concerns about the dog's well-being. But these shifts of attention are all too common in our lives and we do not find them generally threatening to our abilities to keep in mind and respond to other things as well. Alessandra does not seem to be in any way prevented of thinking about Sheba, nor is she deprived of her abilities to do so. Things could have easily happened so (putting aside the consequences of determinism), that at one point of the debate with the school management Alessandra suddenly realizes that Sheba is in the car and rushes out to check her.

So far I was arguing that none of the familiar reasons why we deny one's control over something is able to explain why we would lack control over our negligent behavior. Still, there is something deeply problematic about involuntary omissions: the lack of consciousness. It is not only that we do not *choose* to forget or not notice certain things—we aren't even aware of our wrongdoing at the time of its happening. Forgetting, not keeping in mind, not noticing and not paying attention essentially involve the lack of awareness of certain facts, considerations or reasons. But how could we control something if we aren't even aware of it?

This clash of intuitions leads us to a fascinating issue about the relevant form of control in question. On the one hand, negligent behavior does not involve any kind of disability on the agent's part, but, on the other hand, she is still unaware of the wrongdoing she commits. This raises the question whether the Control Principle requires the actual execution of active and conscious control or the ability to exercise control of this kind. Whereas the former seems to exclude instances of involuntary omissions, the latter can arguably handle these cases. It is remarkable how successful, for instance, Moorean compatibilists are in explaining how agents can possess free will and yet be engaged in negligent behavior. Here is a recent proposal of Ferenc Huoranszki, which touches upon the distinction under discussion:

> But even if it is true that whenever, for instance, we forget to perform an action we cannot intentionally control whether or not we perform it, our will can be free and it is this that grounds our responsibility. What explains our responsibility is that even if we cannot always voluntarily control whether or not we exercise our ability of choice we do not thereby lose the ability to make the relevant choices. Hence, what matters for responsibility is the possession of certain powers and abilities even when we fail to exercise them. Our freedom of will as a condition of responsibility requires *the ability to make a choice about the performance of a kind of action* but it does not demand that we actually intentionally control—either directly or indirectly—what we actually do. (Huoranszki 2011, p. 47)

According to this account of free will and responsibility, Alessandra is responsible for failing to rescue Sheba from the car, because at the time of her failure she had the ability to choose to rescue Sheba and consequently could have saved Sheba, if she had chosen to do so. Without exploring Huoranszki's proposal in details, it is clear that the central move is to identify free will by an (often) unexercised ability instead of exercised conscious control.

But how plausible is it to characterize the relevant control condition in terms of abilities? Here intuitions about ordinary cases become fader and less reliable. At the first glance examples seem to underline Huoranszki's point. For instance, I have control over the windscreen wiper of my car (given that it functions properly, I know how to use it and my hands are not paralyzed) even though it doesn't even come to my mind to use it. I certainly *have* control over it and *can* control it—but do I *control* it after all? And do I control the car itself when it is parking in the garage and I have no thought whatsoever of using it? It would be a little odd to say so. But it would be even stranger to contend that I *cannot* control it or that it is *beyond* my control, just because I'm not using it at the moment. It would be harsh to rely solely on linguistic intuitions here. But maybe it isn't incautious to advance the conclusion that although we might be hesitant to attribute actual control to someone who isn't at the moment consciously engaged in controlling something, it makes a good sense to say that she is still able to control it. Moreover, when we deny someone's control, we do deny it in this second sense: referring to the lack of control as an excuse or exemption amounts to denying the *possibility* of exercising control (at least at the given moment), not only to the acknowledgement that she didn't exercise conscious control (at the given moment). But if, according to the Control Principle, holding people responsible is unfair only if they lack the *ability* to control (either temporarily or in principle), then responsibility for negligence is compatible with the principle.

In this chapter I tried to explore our ordinary concept of control and how we use it for excusing or exempting someone from responsibility. I suggested that although our control over our actions might be the most perfect example of control execution, it might be a mistake to use it as a paradigm example. Instead, I examined the conditions under which we affirm or deny someone's control over external objects, and then I confronted these observations with so-called problematic cases of responsibility, i.e., attitudes and negligent behavior. I argued that given our everyday intuitions about control, we have no good reason to stick to the claim that our judgments of responsibility for attitudes and negligent behavior go against the Control Principle. This obviously does not mean that now we have an account which would explain and justify responsibility-attributions in these cases. First, I did not offer an alternative concept of control. And second, there might well be other considerations than the Control Principle which limit the scope of moral responsibility. However, given what has been said so far we can conclude that appreciating the Control Principle does not entail that we should give up holding people responsible for attitudes and negligent behavior.

Chapter 7: The Normative Interpretation

In Chapter 2 I argued that the concept of control plays a central role in discussions of responsibility because of the Control Principle, which states that it is unfair to hold people responsible for what is beyond their control. I also suggested that the Control Principle led many authors to suppose that control is a necessary condition of responsibility. By having said so I presupposed that the conditions of responsibility and the conditions under which it is appropriate, more precisely, fair to hold people responsible are intimately linked. The inconsistent triad that I presented in the previous chapter contains a hidden premise. More accurately, the argument generated by the Control Principle would go like that:

> P1: It is unfair to hold people responsible for things over which they do not exercise control.
>
> P2: In problematic cases of responsibility the agents do not exercise control.
>
> P3: One is morally responsible for something, only if it is fair to hold her responsible for that thing.
>
> P4 (from P1 and P2): In problematic cases it is unfair to hold agents responsible.
>
> P5 (from P3 and P4): In problematic cases agents are not responsible.

Ordinary judgments of responsibility, as we have seen, tell against this conclusion, since we intuitively find the agents responsible in the problematic cases. In the previous chapter I explored a line of argument which casts doubt on P2 and thus opens up the way of holding people responsible in the problematic cases without violating the Control Principle.

The assumption expressed by P3, just as the Control Principle itself, remains implicit in many discussions. Although for the first glance the premise seems unproblematic, it raises difficult and much debated issues about the concept of responsibility and its relation to responsibility-attributing practices. In the following I would like to explore the supposed connection between being responsible and holding responsible and argue for the acceptance of P3.

7.1 Strawsonian Compatibilism and R. J. Wallace's Account

The most detailed and careful analysis of the topic has been provided by R. Jay Wallace in his influential work *Responsibility and the Moral Sentiments* (1995). Wallace undertakes a monstrous project in his book: by disambiguating and elaborating on the ideas and insights of Peter Strawson's "Freedom and Resentment" he develops a full-blown, coherent Strawsonian theory of moral responsibility.

As I briefly mentioned in the introductory chapter, Strawsonian compatibilism is still a prominent trend in the contemporary literature on free will and moral responsibility. Strawsonianism, however, is a notoriously ambiguous category. In "Freedom and Resentment" Strawson put forward several theses and arguments, the relationship of which are still often discussed and debated (see e.g. Szigeti 2012). Strawson's position on certain crucial issues is undetermined; some of his claims are not supported by any argument, or, more often, they are supported by incompatible arguments (see e.g. Russell 1992). Accordingly, Strawsonians might defend radically different theories, depending on which particular thesis they are committed to.

I find it reasonable to say that Strawsonianism most often involves a special emphasis on moral sentiments in the description and explanation of responsibility-attributing practices. It is hard to say anything more accurate, since positions show a great variation, but probably the most well-known element of the Strawsonian account of moral responsibility is the idea that without making reference to so-called reactive attitudes and sentiments it is impossible to provide an adequate picture of our practices of holding responsible. So, for instance, those who insist that moral blame, the paradigm form of holding someone responsible, necessarily involves affective reactions take a Strawsonian approach in this sense.

However, the Strawsonian conception of responsibility brought novelty also in other respects. What admirers and followers of Strawson have found the most ingenious and perplexing in his account is the way he tied up the concepts of responsibility, reactive attitudes and the issue of justification. Here are the first sentences of Gary Watson's summary of "Freedom and Resentment":

> As his title suggests, Strawson's focus is on such attitudes and responses as gratitude and resentment, indignation, approbation, guilt, shame, (some kinds of) pride, hurt feeling, (asking and giving) forgiveness, and (some kinds of) love. All traditional theories of moral responsibility acknowledge connections between these attitudes and holding one another responsible. What is original to Strawson is the way in which they are linked. Whereas traditional views have taken these attitudes to be secondary to seeing others as responsible, to be practical corollaries or emotional side effects of some independently comprehensible belief in responsibility, Strawson's radical claim is that these "reactive attitudes" (…) are *constitutive* of moral responsibility; to regard oneself or another as responsible just is the proneness to react to them in these kinds of ways under certain conditions. Watson 1987, p. 120)

Then later:

> The explanatory order is the other way around: It is not that we hold people responsible because they are responsible; rather, the idea (*our* idea) that we are responsible is to be understood by the practice. (p. 121)

Following Manuel Vargas's terminology we can say that Strawson's account is *practice-based*. He holds that facts about responsibility are not antecedently fixed; rather, "the 'truth maker' for claims about responsibility is some normative feature of responsibility-characteristic practices" (Vargas 2004, p. 225).

Soon I will say much more about how we should understand practice-based accounts and in particular about Wallace's interpretation. Now it is important to note that this aspect of Strawson's theory, together with Strawsonian naturalism, led to Strawson's much discussed and debated claims about the impossibility of „external justification" of responsibility-attributing practices in general. This is yet another feature of Strawsonianism which divides authors generally sympathetic to Strawsonian ideas.

What I would like to emphasize is that one can commit to any of these elements of Strawsonian thinking without being automatically committed to any other of his claims. As I pointed out in the introductory chapter, only few authors accept Strawson's „no need for justification" argument (or arguments), whereas probably the majority of contemporary authors would admit the special significance reactive attitudes and sentiments play in responsibility-attributing practices. But, as Watson rightly claims, this is far from subscribing to the practice-based approach of moral responsibility: most philosophers would deny that reactive sentiments had a constitutive role in establishing facts about responsibility.

As I see it, Wallace himself accepts both the practice-based approach and (with major restrictions) Strawson's characterization of the relevant responsibility-attributing practices, but refutes Strawson's claims about the irrelevance and impossibility of external justification.[13] In the following I would like to explore only the practice-based aspect of Wallace's account – what he calls the normative interpretation. My defense of Wallace, as it will soon become clear, rests on the assumption that one can accept the normative interpretation without embracing Strawson's characterization of the relevant practices of holding responsible. I would maintain that, despite giving up the special emphasis Strawson

[13] At least it is reasonable to so infer, given that Wallace puts forward several substantive arguments in order to challange incompatibilism.

lays on reactive attitudes, such an account would still count as Strawsonian—but this is mostly a verbal dispute anyway.

In *Responsibility and Moral Sentiments* Wallace argues that determining the conditions of moral responsibility is an essentially normative project: we are searching for those facts which make holding someone morally responsible *fair*. Wallace offers—what he calls—a normative interpretation of what it is to be responsible for something. According to Wallace

> (*N*) *S* is morally responsible (for action *x*) if and only if it would be appropriate to hold *s* morally responsible (for action *x*). (p. 91)

Wallace understands the moral appropriateness in question in terms of fairness, so, following Manuel Vargas (2004), it might be helpful to work with a modified version of *N*:

> (N') S is morally responsible (for action x) if and only if it would be fair to hold s morally responsible (for action x).

It is important to note that by providing *N'* Wallace aims to offer a general schema in order to understand the debate between different theories of moral responsibility, rather than a substantive theory of moral responsibility itself. So as to develop it into a theory, we need to define and characterize both the practices of holding people responsible and the relevant notion of fairness applying to these practices.[14]

By endorsing *N'* Wallace opposes to two rival interpretations: in his terminology, the metaphysical and the extreme pragmatist ones. Whereas metaphysical interpretations (which are far the most common) claim that facts of responsibility are conceptually prior and

[14] There might be a third variable as well, which Wallace does not take into consideration, i.e., the extension of *x*. Wallace assumes that we can be morally responsible only for actions and omissions. However, as we have seen, we have good reasons to suppose that we can be responsible also for our attitudes. At the end of the chapter I will return to this point.

independent of practices of holding people responsible, extreme pragmatist approaches give up altogether the idea of there being any facts about being responsible. The normative interpretation, according to Wallace, offers an intermediate strategy by admitting that there are facts about being responsible, while at the same time he takes them to be conceptually dependent on practices of holding responsible.

Before considering the relative merits of the normative interpretation over the metaphysical one, it is important to make it clear what the normative interpretation exactly claims. The biconditional proposed by Wallace is the conjunction of two conditional statements: that if holding someone responsible is fair, then the agent is responsible and that if holding someone responsible is unfair, then the agent is not responsible. Obviously, both conditionals are necessary if we want to claim that facts about responsibility are defined by the conditions under which it is fair to hold someone responsible. Consequently, Wallace's view can be attacked by refuting one or both of the statements. Interestingly, all the objections raised against the normative interpretation deny the second conditional, ie., that if it is unfair to hold the agent responsible, then she is not responsible. This was exactly the hidden assumption which I used in the last chapter: by accepting that it is unfair to *hold someone responsible* for things beyond her control I inferred that control is a necessary condition of *being responsible*. However, if critics denying this conditional are right, then the Control Principle does not force us to conclude that control is a necessary condition of responsibility. Since this finding might affect to a significant degree all the theories making use of the concept of control, it is worthwhile to examine some recent criticisms.

7.2 Critics I – Angela Smith

Criticisms against Wallace's normative interpretation point out the counterintuitive consequences of his account. To put it in general terms, the normative interpretation cannot make room for those cases where judgments about responsibility and questions about the proper moral response come apart. Whereas the metaphysical interpretation left it entirely undetermined what the appropriate moral response should be, according to *N'* facts about responsibility are to be determined by exactly these responses. As Gary Watson puts it: "In a Strawsonian view, there is no room for a wedge between the practices that evince the reactive attitudes and the belief in responsibility." (Watson 1987, p. 283) Meeting the conditions of moral responsibility, according to Wallace's account, settles decisively what the moral judge should feel and do.[15]

But this is all too often not the case. In her paper "On Being and Holding Responsible" (2007) Angela Smith aims to defend the metaphysical interpretation of moral responsibility against Wallace's theory. To establish her conclusion, she distinguishes three types of moral considerations which bear on the appropriateness of blaming, while being independent of judgments of responsibility. However, if considerations about the fair moral response are, at least partly, independent of questions about one's responsibility, then *N'* is obviously false.

Smith starts out by claiming that

> There seems to be fairly general agreement […] over what is involved in saying that an agent is morally responsible for some thing: to say that a person is morally responsible for some thing is to say that it can be attributed to her in the way that is required in order for it to be a basis for moral appraisal (where nothing is implied about what that appraisal, if any, should be). (Smith 2007, pp. 467–468)

The next step she takes is that she identifies "holding responsible" with blame, where blame is characterized, roughly in line with Wallace's theory, by the expression of reactive attitudes

[15] Both Strawson and Wallace make room for cases where, for some reason or another, we do not actually feel or express reactive attitudes, even though it would be appropriate to do so (since the conditions of responsibility are satisfied). Still, both of them maintain that it is not only the case that it would be fair to feel and express these attitudes – *we are disposed* to feel and express them, if conditions of responsibility are met. For a recent defense of this claim see Wallace 2011.

and sanctioning behavior. At the end Smith presents those three types of considerations which influence the appropriateness of blame, but not the agent's responsibility: the moral standing of the blamer, the significance or seriousness of the fault and the agent's response to her own wrongdoing.

It is difficult to estimate the power of Smith's argument in rejecting the normative interpretation. First and most importantly, proponents of Wallace can argue that Smith missed the point at her very first step. The "fairly general agreement" to which Smith refers when offering her definition of being morally responsible is an agreement between supporters of the metaphysical interpretation, the interpretation to which Wallace oppose. We cannot reject the normative interpretation on the basis that if we accept a different interpretation, then the normative interpretation will not work. Obviously, if being responsible for something means openness to moral appraisal, then, at the same time, it won't be true that being responsible is being a fair target of blame. However, this does not show that the latter understanding is inferior to the former. Proponents of Wallace can consistently go on to say that, according to the normative interpretation, agents in Smith's examples are not morally responsible, since it is unfair to hold them responsible.

Still, these examples provide good grounds to think that the normative interpretation is utterly counterintuitive. How can Wallace explain why the considerations which Smith mentions seem external to issues of responsibility? How can he explain away our intuition that agents can be morally responsible even when any kind of blaming behavior would be inappropriate toward them?

I find two strategies open to the supporters of Wallace, only one of which I will discuss in details. The first strategy is to argue that it is not the *fairness,* but other, ethical dimensions of blame which can be questioned among the lines of these considerations. It can be quite unkind, unnecessary, ungenerous, uncharitable, or even mean to blame someone in certain situations. However, this does not have any bearing on whether it is fair to hold these people responsible, given the relevant notion of fairness. This strategy obviously requires a specific concept of fairness that is not reducible to other kinds of moral evaluations.

The second, more ambitious line of reasoning would be to accommodate Smith's aforementioned definition of being responsible within the normative interpretation. Proponents of Wallace can claim that the supposed differences in attributing responsibility between Wallace's and Smith's accounts arise not because Smith endorses the metaphysical, while Wallace the normative interpretation, but because they characterize responsibility-attributing practices in a radically different way. At this point we need to remember that the normative interpretation is not a theory, but only a schema, whose variables, i.e., the relevant form of holding responsible and the concept of fairness, can be filled in in many different ways. Whereas Wallace takes holding responsible to be "a highly structured set of emotions and actions" (p. 88), for Smith the relevant practice attached to responsibility is the formation of judgments about the agent's actions and attitudes.[16] The reason why considerations of fairness seem unnecessary and external for Smith in establishing one's responsibility is that she characterizes holding responsible by such "milder" practices which, in contrast to reactive attitudes and sanctioning behavior, do not raise fairness issues. If holding responsible is nothing more than forming a judgment about one's moral qualities, then it will be always fair if the agent (her actions or attitudes) in fact has the relevant qualities. The fairness of judgments, especially if they are never expressed, depends only on whether they are true or false. Reactive attitudes and sanctioning behavior, by contrast, might be morally inappropriate for other reasons, some of which were listed by Smith.

[16] This becomes more evident in Smith (2005, 2008). In the next chapter I will discuss her notion of moral criticism in length.

7.3 Critics II – Dana Nelkin

Another, similar objection was raised by Dana Nelkin (2009). The target of Nelkin's criticism is not the normative interpretation, but what she takes to be the most powerful incompatibilist argument, the *The Intrapersonal Fairness Argument Concerning Blameworthy Actions*:

> P1: *X* is responsible and blameworthy in the accountability sense for an action *a* only if it would be fair to impose sanctions on *X* for *a*.
>
> P2: It would be unfair to impose sanctions on *X* for the performance of an action *a* if *X* lacked the ability to do otherwise.

Therefore,

> C: *X* is responsible and blameworthy in the accountability sense for the performance of an action *a* only if *X* had the ability to do otherwise. (Nelkin 2009, p. 152)

Nelkin attacks P1 by considering historical figures like Mahatma Gandhi, who advertised that injustice should be cured without sanctions and reactive attitudes. While Gandhi was aware of the wrongs to be repaired and fought against, he disapproved sanctions and retributive sentiments toward those who are responsible for them. According to Nelkin,

> the possibility of Gandhi (or someone like him in this respect) suggests that it is not incoherent to say that one is obligated to meet certain standards without being automatically committed to the claim that it is fair to impose sanctions for failing to meet them. If the link between accountability and the fairness of sanctions *could* be broken in this way – so that accountability did not *by itself* entail fairness of sanctions – then the claim that avoidability is required for fairness of sanctions could simply fail to apply to its target. (Nelkin 2009, p. 157)

Here I won't present Wallace's own answer to Ghandi-like cases, because it is largely based on his specific understanding of what holding responsible involves. Nevertheless, similar points as those which I proposed concerning Smith's objection can also be raised here. First, as I've previously emphasized, holding responsible need not involve sanctions and reactive attitudes. It can be characterized this way—as a matter of fact it is a quite typical way of characterizing it, but that does not tell in itself against the normative interpretation, put forward as a general schema.

Second, I find the step between saying that it is *fair* to hold someone responsible (whatever that may involve) and affirming that we *should* all things considered hold her responsible, too quick. I find two viable ways to block this inference. On the one hand, we can make a good sense of talking about the fairness of doing something without supposing that we are *obliged* to do so or even that it would be a *morally superior* state of affairs if we did so. Sometimes, by stating that something would be fair (or not unfair) we simply state that the given course of action is morally *permissible*, insofar as fairness is concerned. Suppose that a friend of mine was asked by another friend to give a hand in her moving to a new apartment. My friend, however, is quite busy and not really in the mood of lifting boxes all day. If she asks me whether it would be morally wrong to stay at home instead and do her own task, I may well reply: "No, it wouldn't be unfair, last year she didn't help you either". In this case I obviously don't want to imply that she *should* stay at home, even less that it would be *unfair* or morally inappropriate to go and help her friend. I simply affirm that by staying home she does not violate any duty of fairness. I find that we have good reasons to assume that we use fairness in a similar vein when we talk about the fairness of holding someone responsible. Interestingly, our intuitions about when it is unfair to hold someone responsible are much stronger than those positively supporting the fairness of holding responsible. This observation might well indicate that when determining the fairness of holding someone responsible in a particular case we are primarily interested in the permissibility of engaging in this practice.

But even if we insist that considerations about fairness always create obligations, we have no reason to suppose that these considerations are always *overriding*. As I can see it, Gandhi's reluctance toward sanctions was part of a larger ethical framework of his in which peace and cooperation played a central role. One can assent to the claim that certain sanctions under certain circumstances are fair, while claiming that considerations of fairness in the given, still unspecified sense are *in general* less valuable than competing moral reasons which tell against sanctioning and the expression of reactive attitudes. I find this position not only coherent, but also historically more accurate than the one proposed by Nelkin.

7.4 Normative versus Metaphysical

So far I have argued that the apparently counterintuitive consequences and counterexamples can be eliminated, if we take it seriously that the normative interpretation is not a theory but a schema which can be developed in many different ways. But I did not offer any positive reason to accept the normative interpretation over its rivals, the extreme pragmatist and the metaphysical interpretation. Since extreme pragmatist accounts are both rare[17] and notoriously troublesome, here I will focus on Wallace's remarks only about the metaphysical interpretation. He writes,

> the practice of holding people responsible is characterized by a highly structured set of emotions and actions, namely the reactive emotions and the blaming and sanctioning behavior that expresses them. But it seems incredible to suppose that there is a prior and independent realm of facts about responsibility to which such emotions and actions should have to answer. (…) Admittedly I have not myself shown that this picture could not have an application. I think the burden is on someone who wishes to interpret the issue in these metaphysical terms to defend and develop the supposition that there is a prior and independent realm of facts about responsibility—something I, for one, cannot see how to do. (Wallace 1995, p. 88)

[17] Wallace refers to Honderich (1986) and I couldn't find any other clear example for such view.

Wallace's criticism is somewhat puzzling and, as he himself admits, hardly amounts to a conclusive argument for the superiority of the normative interpretation over the metaphysical one. Given what has been said so far, proponents of the metaphysical interpretation can consistently go on to assume that facts about responsibility are prior and independent of responsibility-attributing practices, but have significant normative *consequences*. That is, they ensure that the agent is open to some kind of moral appraisal on the basis of her action. Even if we explain away all the possible counterexamples similar to those presented by Smith and Nelkin, defenders of the metaphysical interpretation can still insist that Wallace simply reverse the order of explanation: it is not the fairness of holding responsible which fixes facts about moral responsibility but, on the contrary, it is fair to hold people responsible because they *are* responsible.

Surely, their position is not the least trivial: they owe us an explanation of how it happens that the fairness of seemingly unconnected practices such as moral criticism, blame and sanctioning behavior track perfectly reliably these metaphysical facts of responsibility—most probably in the quote above Wallace refers to this problem. But even if the burden of explanation is on the proponents of the metaphysical interpretation, this in itself does not show that they are wrong. Moreover, admittedly their construal of the problem seems to be much closer to the ordinary understanding of the relationship between being and holding responsible. As Manuel Vargas (2004) frames the problem:

> As agent-based theorists see it, there are only two things we need to know to learn the facts about responsibility in any particular case: what kind of agent is involved, and the agent's connection to the considered action (or state of affairs). Wallace-style Strawsonians, however, maintain that we need to know a further thing: whether deployment of responsibility characteristic practices is appropriate (or fair, etc.) in general and in the particular case. But, what evidence could they offer for thinking that we need to know these things as well? (…) It seems gratuitous to

> insist that the normative property of being responsible is parasitic on a further, more basic normative property (e.g., the fairness of the practices), which is itself dependent on properties of agency and action on which the status of being responsible was initially thought to depend. (Vargas 2004, p. 225)

I find Vargas's interpretation of Wallace's view uncharitable. Defenders of the normative interpretation do not ask a *further* question about the moral appropriateness of responsibility-attributing practices: this is the *only* question they ask. However, Vargas is right to point out that at the end of the day defenders of the normative interpretation (or at least Wallace) will come up with more or less the same ideas about the conditions of responsibility as supporters of the normative interpretation. But if this is so, then why is the fairness of holding responsible important in the first place?

I find that the only way to answer this objection on the part of Wallace is to argue that supporters of the metaphysical interpretation are also committed to taking into consideration considerations of fairness, even if they are not aware of it. Of course, this claim shouldn't take the form of some awful false consciousness argument: the only thing we need is to show that the task which the metaphysical interpretation aims to accomplish cannot be done without taking normative considerations into account. And this conclusion might come along—here I only sketch a possible argument supporting it.

Advocates of the normative interpretation can legitimately ask how the supporter of the metaphysical interpretation arrived at her conclusions about the conditions of being morally responsible. Most probably the answer would be something along these lines: "I examined paradigm cases of someone being responsible for something and also typical cases when we exempt someone from responsibility. Then I searched for properties which are instantiated at the former, while absent in the latter cases. The presence of these properties (or property) is the necessary and sufficient condition of moral responsibility."

However, this methodology invites unwelcome consequences. Imagine that neuroscientists identify a neurological event in the brain which occurs every time when performing typically responsible instances of human agency, while absent in all those cases when we usually and uncontroversially exempt agents from responsibility. Can this neurological event be the necessary and sufficient condition of moral responsibility?

Setting aside how improbable such a discovery may be, most of us would still say no.[18] There are some commonly accepted restrictions on the kinds of terms in which conditions of responsibility can be given. These are exclusively psychological terms (even if their existence might be dependent on metaphysical facts), mostly referring to, among other things, reasons, desires, choices or the will. Why do we maintain such restrictions? At this point the proponent of the normative interpretation can press the claim that the reason underlying the restrictions is that we require more work to be done by the conditions of responsibility than merely picking out intuitively responsible instances of agency. We also want these conditions to have justificatory force—that is, they have to be able to justify our practice of holding people responsible[19]. By using the strategy sketched above, Wallace would be able to convince proponents of the metaphysical interpretation that facts about responsibility cannot be prior to responsibility-attributing practices, since the whole quest for the conditions of moral responsibility is essentially governed by the requirement of justifying these very practices.

[18] But not Manuel Vargas (2004), who writes: "It could well turn out that our ordinary concept of responsibility has metaphysical commitments that are not normatively justified. But, that would be a discovery about our concept, not something we should rule out as a matter of principle." (p. 226). Here I would only like to note that accepting Vargas's claim leads to the denial not only of the normative interpretation, but also the weaker claim, endorsed by virtually any proponent of the metaphysical interpretation, i.e., that facts of responsibility have normative consequences.

[19] A similar objection has been put forward by Dennett (1984).

7.5 Critics III – John Fischer

Can supporters of the metaphysical interpretation resist the conclusion? Another recent criticism of Wallace, put forward by John Fischer (2011), might further illuminate the nature of the debate. Fischer explores the metaphysical-normative distinction, as presented by Gary Watson, Susan Wolf and R. J. Wallace. Although Fischer's final conclusion is that each author understands the supposed normativity of her approach somewhat differently, he returns again and again to the idea of *non-reductionism*. First he gives the following interpretation of Wallace's normative interpretation:

> The point is presumably that we cannot replace the right side of the biconditional [the biconditional is what I have called *N'* – A. R.] above with any purely descriptive condition (any condition in which normativity does not play a certain distinctive role). (…) Of course, all normative approaches would accept the claim of non-reductionism; Wallaces specific contribution is the analysis of being responsible in terms of the fairness of holding responsible. (Fischer 2012, p. 138)

Fischer's proposal is that, similarly to Watson and Wolf, Wallace wants to exclude the possibility of specifying the conditions of responsibility in non-normative terms. If this is the correct interpretation of Wallace, then Fischer is perfectly right to complain that he provides no argument whatsoever to accept this claim. Also, I agree that we have no apparent reason to embrace such a restriction prior to engaging in the project of determining the conditions of responsibility. However, after careful examination of Wallace's text I haven't found any clue which would support the claim that he argues for such a restriction. On the contrary, it seems that he takes very seriously the challenge coming from those incompatibilist accounts, which explicate the condition of responsibility in purely descriptive terms. This suggests that we should interpret Wallace in different fashion. Fischer also recognizes this tension and at by the end of his analysis he suggests a different understanding of the normative interpretation:

> Finally, someone might insist that Wallace's approach is fundamentally different (say) from mine in giving analytic hegemony to the (explicitly normative) notion of "fairness." Wallace, it might be thought, is fundamentally interested in the normative justification—the fairness—of our responsibility practices. Now this might be true, but how exactly is it diff erent from my approach (or those of others who write about moral responsibility—even such "metaphysicians" as Kane, van Inwagen, and O'Connor)? I take it that we are *all* fundamentally interested in the normative issue of whether our responsibility practices are *justified* in light of skeptical worries issuing from both causal determinism and various forms of indeterminism. (Fischer 2012a, p. 141)

And then, somewhat later, slightly indignantly:

> How is this project—which has motivated me in *all* of my work on the various facets of free will (even the traditional problem of the relationship between God's omniscience and human freedom)—different from Wallace's? Similarly, I take it that libertarians offer their (various) accounts of free will precisely because they do *not* think that our responsibility practices would be *normatively defensible* or *fair*, given causal determinism. I believe that Wallace himself would agree with this conceptualization of the dialectic; that is, I think that Wallace would agree that we *all* are interested in the normative foundations of our responsibility practices, but we come to different conclusions about the conditions for the normative defensibility of these practices.

Fischer thus argues that at the end of the day Wallace's main contention is either implausible (and not supported by any argument) or banal, because everyone working on the field of free will and moral responsibility would subscribe to his claim. As opposed to the first interpretation, I find this latter one, including Fischer's diagnosis, entirely correct. Although in some passages Wallace seems to present these two camps as genuine rivals, in the previous paragraphs I tried to make it clear that I do not find the metaphysical interpretation a real alternative, because any enquiry into the conditions of moral responsibility necessarily aims at justifying those practices which we usually associate with holding responsible. However, this does not mean neither that there are no facts about responsibility, nor that these facts are non-reducibly normative.

7.6 Methodological Consequences I – Individualistic and Externalistic Approaches

So far I have been arguing that we have good reasons to embrace (and no good reasons to refute) the normative interpretation, as provided by Wallace. According to this understanding of the concept of responsibility facts about responsibility are defined by the conditions under which it is fair to hold someone morally responsible. I find that this conclusion makes a great help in understanding why we find it all too often impossible to make justice between competing theories of responsibility. Since the normative significance of conditions of responsibility in justifying the fairness of holding someone responsible remains hidden in many theories, their basic assumptions gain a peculiar bedrock status, which we cannot analyze or further reduce by any means.

But why are we defending the normative interpretation, when it just turned out that no one attacks it? Moreover, why is this an important issue, if after all in this respect there is no real controversy between the different parties? My short answer is that because the normative interpretation, correctly understood, is not a theory but a schema, its significance lies not in a particular claim, on the basis of which theorists of the field can be divided into groups, but in its methodological consequences. In the remainder of this chapter I will explore both the limitations and the opportunities which follow from the appreciation of the normative interpretation. Some of these might be obvious, while others more controversial, but it is fair to say that these methodological considerations are quite frequently ignored in the relevant literature.

A novel and interesting methodological question was raised by Andrew Sneddon (2005). Sneddon argues that Strawson's most radical statement about moral responsibility in "Freedom and Resentment" is still not appreciated and even less incorporated into recent discussions about moral responsibility. He interprets the practice-based approach of Strawson's account as arguing for the necessity of acquiring social competences in order to be a responsible agent:

> According to Strawson, to hold someone responsible is to deploy the reactive attitudes towards that person. (…) Presumably then, in the first place, to be responsible is to be an apt candidate for the reactive attitudes. Put another way, to be morally responsible is to fit into the social practices governing the deployment of the reactive attitudes. In short, it is to acquire a *social competence*. (Sneddon 2005, p. 241)

The interesting question is how to construe this social competence. And here comes Sneddon's crucial distinction between *individualistic* and *externalist* approaches of moral responsibility. Whereas individualistic approaches construe the central condition (in this case, social competence) of responsibility exclusively in terms of the agent's intrinsic properties, externalist approaches also allow for relational properties in order to explicate the given condition. Sneddon illuminates the distinction in the following way:

> Here is one important difference between individualistic and externalistic approaches to moral responsibility: if moral responsibility is a competence$_I$, then if one imaginatively holds the intrinsic properties of a morally responsible individual constant but varies the properties of his/her environment, especially the social properties, then the individual will always be morally responsible. (…) By contrast, if moral responsibility is a competence$_E$, then performing the same sort of thought experiment can result in changes to the agent's responsibility. (…) Moreover, if one holds environmental properties constant, then two agents with very different intrinsic properties could both be morally responsible. In the case of at least some externalistic competences, *no* particular intrinsic properties of an agent may be

> necessary or sufficient for their realization. Moral responsibility might be such a competence$_E$. (Sneddon 2005, p. 242)

Although the general distinction between intrinsic and relational properties might be familiar to most philosophers, its application to theories of responsibility is novel and fascinating. Sneddon claims that virtually all accounts of free will and moral responsibility—even such devoted followers of Strawson as Wallace—pursue the individualistic approach. Nevertheless, deeper appreciation of "Freedom and Resentment" suggests that it would be a "natural development" (ibid., p. 239) of Strawson's position to construe the conditions of responsibility in an externalistic fashion.

I do not want to analyze Sneddon's interpretation of Strawson. I am not quite sure that his ideas about responsibility as an externalistically construed social competence would follow so naturally from Strawson's original text. However, this does not make the distinction between individualistic and externalistic approaches any less interesting or important. Also, I find that the introduced opposition is closely related to the supposed difference between the metaphysical and normative interpretation.

Of course, someone could say, in the spirit of Fischer, that this is just another useless way to make sense of a bad distinction. After all, why would an account be any more metaphysical just because it does not allow for relational properties in specifying the conditions of responsibility? (Come on, the distinction itself comes from metaphysics!) And what makes us think that someone cares more about normative justification just because she takes an externalistic approach?

Obviously, I do not want to argue for a one-to-one correspondence between these groups (R. J. Wallace would be an instant and eminent counterexample). Rather, I would like to suggest that appreciating the normative interpretation makes us more disposed to recognize the possibility of an externalistic approach. As long as one sees the primarily task of theories of responsibility in identifying a neat, uniform metaphysical condition which fits to our ordinary judgments of responsibility, one will be prone to think that relational properties (unless the relation is between the agent and her action) are not the kind of things one should look for. If, by contrast, one takes seriously the thought that her task is to provide moral justification for our responsibility-attributing practices, it might turn out that previously neglected, seemingly external factors have a significant role in establishing facts about responsibility.

I have found only one clean-cut example of externalistic approaches, but it is perfect illustration of my previous points. In his recent book, *Justice for Hedgehogs* (2011) Ronald Dworkin argues for a capacity view of moral responsibility. Dworkin's methodological claims show startling resemblance with Wallace's normative interpretation. He makes a sharp distinction between so-called "scientific issues", concerning the effect of causal determinism and indeterminism on human though and action and issues of "judgmental responsibility", which concerns the moral appropriateness of praise and blame on the basis of the agent's behavior. Dworkin claims that

> It is crucial now to notice the large logical space between the first set of issues— the scientific or metaphysical questions that can be answered, if at all, only through empirical investigation or philosophical speculation— and the last set, about responsibility, which are independent ethical and moral issues. (…) [N]o conclusion about responsibility can follow directly from any answers we give to questions in the first set. Any inference from the first to the third set of issues requires a further evaluative premise. The literature of the free will problem has not, in my view, paid sufficient attention to this requirement— perhaps because philosophers assume that it is obvious which ethical and moral principles are

> available to bridge the gap. I believe that this is very far from obvious. (Dworkin 2011, pp. 221–222)[20]

Accordingly, Dworkin regards the quest for the conditions of judgmental responsibility as a task of ethical interpretation, where we explore which moral and ethical principles should guide the practice of moral appraisal in order to maximally fit into our larger ethical framework. According to his account the relevant form of control necessary for moral responsibility should be understood as a capacity to (i) form true beliefs about the world and (ii) match the decisions to the agent's normative personality, i.e., „his settled desires, ambitions, and convictions" (Dworkin 2011, p. 228).[21]

Dworkin discusses the typical excusing and exempting conditions, identifies the ethical principles behind them and explains how his capacity view incorporates these considerations. At the end of the chapter, however, he takes on the challenge of discussing some particularly puzzling cases of moral responsibility, i.e., acting under duress and the mitigation of responsibility in the cases of those who were raised in poverty. Here Dworkin makes a proposal which is independent of his earlier presented capacity view. He suggests that

[20] The second set involves considerations about freedom. These, however, according to Dworkin, collapse back into either the scientific or the responsibility question:

> There is no pertinent question about whether people are free that is not either the scientific or the ethical question in disguise. Some people use "freedom" simply to mean nondeterminism: people are not really free, they assume, unless determinism is false. Others use the word simply to mean responsibility: they say that people are or are not free when they mean that they are or are not judgmentally responsible for their actions. Neither of these ways of speaking is mistaken: it is not a linguistic mistake to say either that people are not really free because determinism is true or that people are really free, even if determinism is true, when they are subject to no external constraint. But talk of freedom in this context is unhelpful and often sponsors confusion. (Dworkin 2011, p. 222)

I could not agree more.

[21] Interestingly, this characterization also shows close similarities to Wallace's view, which we will further explore in Chapter 9.

> we are tempted to find diminished responsibility in these circumstances because— but only when— *duress or poverty is the product of injustice* [my emphasis]. Our foundational responsibility to live well provides a ground for claiming moral and political rights. (…) We might— or might not— think that these rights should be protected by a further, distinct, responsibility filter in addition to the capacity filters we have been discussing. (…) This distinct, further filter is conceptually available because the root questions for the responsibility system are not metaphysical but ethical and moral; this further filter is controversial for exactly that reason. (Dworkin 2011, pp. 251–252)

I find Dworkin's suggestion truly fascinating, but here I cannot engage in the systematical exploration of his ideas. I would only like to point out two features of his proposal, from which we can draw more general morals with respect to our present issues. First, by making social injustice a mitigating factor Dworkin clearly introduces an externalistic element into his theory. It would be far beyond the scope of this discussion to present Dworkin's ideas on justice in general, but it is clear that being a victim of injustice, according to him, depends crucially on the opportunities and resources *other people* have. No matter how much I know about someone's circumstances, I cannot determine whether it is unjust to her to suffer these circumstances as long as I don't know anything about how she got into this situation and the circumstances of other people. Thus, being the victim of injustice is clearly a relational property according to Snedden's characterization and thus renders Dworkin's account externalistic.

Second, as Dworkin explicitly admits it, this requirement is completely independent from the capacity view which he defends and he does not even attempt to tie them together. This is a somewhat peculiar method, compared to more traditional theories of free will and moral responsibility. Most theories strive to give a unified account—even if they cannot explain all judgments of responsibility exactly alike, they at least try to show how different conditions can be traced back to a common root or can be regarded as variations or expansions of one and the same idea (Fischer's and Ravizza's treatment of responsibility for emotions might be helpful example, see Chapter 2).

I find that this ambition can't be traced back solely to aesthetic or economical reasons. Rather, it is a natural consequence of the thought that defining the conditions of moral responsibility involves the discovering of a metaphysical property to which judgments of responsibility respond. If facts about responsibility are prior and independent of responsibility-attributing practices, as an imagined proponent of the metaphysical interpretation would claim, then it is reasonable to suppose, given that no additional argument is provided, that these facts are relatively simple and metaphysically uniform.

If, by contrast, we accept that facts about responsibility are fixed by our moral and ethical standards regulating the fairness of responsibility-attribution, then we have no reason to suppose that these considerations will, at the end of the day, add up to a uniform theory. As we will see later, several kinds of moral considerations regulate our intuitions about this issue, and it would be unreasonable to assume in advance that one single principle would win out and contribute to a uniform account of responsibility. Appreciating the significance of the normative interpretation might lead to theories of responsibility which are less neat and pretty than the ones we are accustomed to.

7.7 Methodological Consequences II – Extension, Fairness, Moral Practices

In every other respect, however, the normative interpretation sets higher or at least more complicated standards to theories of responsibility. As I have previously mentioned, the normative interpretation is only a schema which can be filled in in many different ways. According to my understanding, this schema contains three variables.

The first one is the *extension* of responsibility: what are those things for which we can in principle responsible? Wallace is silent about this problem since he presumes that we can be responsible only for our actions and omissions. However, as the previous chapter made it clear, our current practices support a much wider extension of responsibility-attribution. Since we often make moral judgments about others' attitudes and involuntary omissions, we have to consider the possibility that our responsibility extends far over our voluntary actions and omissions.

Second, we have to give an account of the concept of *fairness* in the present context. As we will see in Chapter 9,e way we characterize fairness has a significant bearing on the resulting theory. For instance, while many authors (typically incompatibilists) offer desert and merit-based accounts of fairness (see e.g. King 2012, Pereboom 2001), others (mainly compatibilists) are more eager to identify unfairness with unreasonableness (Sher 2005, Wallace 1995).

The third factor to be considered is the nature of the moral practices involved. As long as we do not have a clear picture of what responsibility-attributing practices such as praise, blame and reactive attitudes imply, it is impossible to tell, the fairness of *what* should be guaranteed by the theory. If resentment, on the one hand, involves the thought that the agent has done something wrong and deserves punishment for her action, then our theory should establish a robust notion of desert. But if, on the other hand, resentment is only a reaction to an agent's voluntary action which caused harm, then arguably no stringent metaphysical or psychological conditions has to be met to make resentment morally justified. At the most extreme, as we shall see in the next chapter, one can argue that the conception of responsibility she offers does not involve sanctions and rewards at all. If we provide such a weak notion of moral criticism as a response to responsible agency, then it might be that issues of fairness do not arise at all.

I take it that a theory of responsibility should not only accomplish these tasks in a coherent way, but also has to create a reflective equilibrium between the theory and our ordinary thinking about the extension of responsibility, the concept of fairness and the nature of our moral practices. It might be—and the literature on free will and responsibility makes this claim likely—that this task cannot be completed without being revisionist in at least one aspect of the three. As the following scheme demonstrates it, revisionism can arise in any of the three spheres:

	Extension of Responsibility	Fairness	Responsibility-attributing practices
E.g. hard determinists	revisionist	non-revisionist	non-revisionist
E.g. some compatibilist theories	non-revisionist	revisionist	non-revisionist
E. g. attributionism	non-revisionist	non-revisionist	revisionist

I take it that a theory of responsibility is non-revisionist if:

a) It conforms to our ordinary judgments of responsibility *(extension)*

b) It works with a relatively robust, most probably desert-based concept of fairness, which at least partly reflects our intuitions about retribution *(fairness)*

c) It allows for ordinary responsibility-attributing practices (praise, blame, reactive attitudes) as justified responses to responsible agency and admits their sanctioning/rewarding nature *(responsibility-attributing practices)*

The interrelation among these factors, according to my interpretation, works out in the following way. If we want to save b) and c), then most probably we have to limit radically the scope of responsible agency in order to make sanctions and rewards morally justified. However, if our aim is to save as much of our ordinary judgments of responsibility as we can, then we have to weaken the normative force of holding someone responsible, either by defining fairness in a less robust way (one that probably won't justify punishment, for instance) or by defining the nature of moral criticism so that issues of fairness do not arise at all. Strictly speaking b) and c) are not independent factors: the more severe the sanction involved in responsibility-attribution is, the more robust our concept of fairness has to be in order to justify the practice. However, I still find it useful for the present purposes to distinguish the two strategies, because it touches upon one of the main characteristics of attributionist accounts.

I do not want to claim that every theory of responsibility fits nicely into this mapping. This is mainly because most authors do not formulate the problem of free will and moral responsibility the way I proposed. The problem of fairness remains hidden or implicit in most of the works on the subject and even if it doesn't, almost no one gives a detailed analysis of the normative concepts involved. I think that this deficit can be explained by the inadequate appreciation of the normative interpretation, which comes together with a systematic and purposeful negligence in spelling out the details of practices of holding responsible. Hopefully this chapter might have convinced some that the normative issues will catch them, no matter how hard they try to hide.

7.8 Methodological Consequences III – Fifty Shades of Responsibility

Adopting the normative interpretation, however, has a somewhat peculiar consequence. Given what has been told, it seems that applying this schema will lead us to accept that there can be several adequate theories of responsibility which cannot challenge one another. Depending on how we characterize fairness, responsibility-attributing practices and the extension of responsibility, we will accept (sometimes radically) different conditions of responsibility, which will be suitable only relative to the normative notions we have initially chosen.

First, this idea is not altogether unfamiliar in the literature. Recently, John Fisher and Neil Tognazzini (2011) distinguished 15 (!) analytical stages in examining the concept of responsibility, each of which can raise further considerations about the conditions of responsibility. Fischer and Tognazzini argue that compatibilists and incompatibilists often talk past each other because they are answering different questions about responsibility. As they put it: "It is one thing to talk about the connection the agent has with her action; it is quite another to talk about the potential interaction the agent might have with her moral community" (Fischer & Tognazzini 2011, p. 381).

Second, the normative interpretation, as I see it, does not force us to accept just whatever theory one happened to come up with. There are several methodological and normative restrictions in play which radically limit the possible number of acceptable theories. We need to ask, for instance, whether the given concept of fairness or the characterization of responsibility-attributing practices is really the most relevant one. It can be argued that identifying holding responsible with the formation of moral judgments (in an attributionist fashion) cannot do the job, since the resulting theory won't be able to give an account of more typical and central phenomena such as blame and reactive attitudes (I will explore this criticism in the next chapter). The same can be said about fairness: if someone thinks that the relevant notion of fairness should be a desert-based one, then she will reject any theory of responsibility that works with a different notion of fairness. I do not want to claim that these are settled issues: the only point I want to make is that such considerations limit the number of admissible substantive theories of responsibility even if one takes the normative approach.

Another restriction on the possible set of theories of responsibility comes from the interrelation of the related concepts. The task of determining the conditions of moral responsibility is the achievement of a reflective equilibrium between responsibility-attributing practices, fairness and the scope of responsibility. We have to justify certain practices as responses to other practices, and this process is not the least arbitrary. It would be quite troublesome, for instance, to find such a notion of fairness which would make it fair to punish people for their emotions.[22] This is because justifying the fairness of punishment involves a wide range of normative considerations and emotions are not likely to meet the stringent conditions which result from the requirements of these considerations. Or, the other way round, the milder we take responsibility-attributions to be, the less robust justificatory account is needed to guarantee their fairness. This way more instances of human agency can be regarded as responsible.

This last point about the plurality of responsibility concepts can also explain why some attributionist authors maintain that responsibility has more than one faces. As we have seen, Thomas Scanlon makes a distinction between substantive responsibility and responsibility as attributability, while Gary Watson assumes that there is another face of responsibility, accountability, which has more stringent conditions than attributability. These non-uniform treatments of moral responsibility make sense as soon as we regard them as different ways of to characterize responsibility-attributing practices and thus the moral principles determining

[22] Schlossberger (1986) might disagree.

the fairness of them. But even if attributionists provide a unified account of responsibility (as the one proposed by Smith), we may still ask if this concept of responsibility is the same as the one discussed by other authors. As Michael McKenna and Neil Levy (2009) rightly points out:

> Attributionists often distinguish between two kinds of responses to wrongdoing: holding responsible and blaming (or imposing sanctions). The latter, but not the former, require the satisfaction of control conditions, they argue (…). This fact suggests the possibility that the debate between attributionists and volitionists is as much verbal as substantive: it may be that volitionists identify responsibility with the attributionist's blaming, whereas what attributionists call responsibility the volitionist might call (…) areatic criticism. Making progress on the debate therefore requires arguments for identifying responsibility with one or other notion. (Levy–McKenna 2009, p. 118)

In the following chapter I will explore the discussed notions of holding responsible, moral criticism and blame as provided by attributionist accounts. But before turning to the next topic, let's return briefly to the initial problem about the Control Principle and its possible consequences. As I said, it seems natural to shift from the statement that it is unfair to hold people responsible for things beyond their control to the claim that control is a necessary condition of moral responsibility. But this inference presupposes that being responsible and the fairness of holding responsible come hand in hand, a thesis refuted, among others, by Angela Smith. Although in this chapter I argued that her objections, as well as others questioning the legacy of the normative interpretation, are unsuccessful, it is still worthwhile to pose the question: what would follow from the falsity of it?

Although the Control Principle might yet still be true, it would not be a really interesting or important thesis, since nothing would follow from it. If we accept, following Smith, that the fairness of holding responsible is determined by considerations external to facts about responsibility (and thus refute P3), then we can continue to think that someone is morally responsible even though it would be unfair to hold her responsible. Consequently, even if the lack of control would render responsibility-attributing practices unfair, it would not excuse or exempt the agent from responsibility. Surely, this would be a happy conclusion to Smith, who denies that control matters to responsibility at all and so it motivates her objections against Wallace. It is important to note that this strategy to dissolve the clash between everyday judgments of responsibility and the Control Principle differs from the ones discussed in the previous chapter: instead of revising our concept of control or denying the clash, the form of attributionism presented by Smith simply denies that the principle has any bearing on our judgments of responsibility. I don’t find this strategy viable. However, as we will see soon, other attributionists have further resources to undermine the significance of control.

Chapter 8: Moral Criticism and Blame

Attributionists argue that we can be morally responsible also for what we cannot voluntarily control, most notably for attitudes and negligent behavior. This claim, as we have seen, apparently goes against an appealing moral principle, the Control Principle, which asserts that it is unfair to hold people responsible for what is beyond their control. In chapter "What Do We Mean By Control" I tried to show that we have sufficient grounds to argue that we do have control over our attitudes and negligent behavior, even if it is less perfect than the one we have over our actions. Nevertheless, this is not the strategy which attributionists themselves follow. In the previous chapter I presented another way of dissolving the clash between the Control Principle and attributionism presented by Angela Smith, i.e., denying the inference from the unfairness of *holding responsible* to the negation of *being responsible*. This strategy, were it successful, would amount to showing that although the Control Principle might be valid, it has no bearing whatsoever on the conditions of responsibility. However, as I argued in details, Smith's objections fail; moreover, the debate between attributionist accounts and other theories can be best understood within the framework of the normative interpretation which Smith attacks. The normative interpretation, in turn, welcomes us to consider the form of moral appraisal or blame which these accounts endorse. The attributionist task is now to show that it is *not* unfair to hold people responsible for what they cannot control, and in order to do that first they have to specify their concepts of holding responsible.

Questions can be articulated in two ways concerning the descriptive and normative aspects of blame and moral appraisal. First, we can ask what kinds of facts we presuppose when blaming someone and, consequently, whether these facts really do obtain. When asking this we test the theory's descriptive adequacy. This issue arises most dominantly in the debate between compatibilists and incompatibilists and confronts our preliminary presuppositions with the consequences of the truth of determinism. This way of posing questions about blame has a definitely cognitivist flavor: it assumes that blame is backed up or constituted by certain judgments, which are truth-apt: they can be either true or false. This idea might seem to go against the Strawsonian concept of blame (endorsed, among others, by Wallace), which essentially involves reactive sentiments such as resentment or indignation. But the contrast is apparent, since these emotions can also be modeled in a cognitivist fashion. Moreover, even if we accept a non-cognitivist theory of emotions, we can still sensibly talk about the *fittingness* of certain emotions (see D'Arms & Jacobson 2000), which consideration leads us back to questions about descriptive adequacy.

But there is a second question, which we already encountered in the last chapter: under what conditions is it *fair* to blame people for their actions and attitudes? As we have seen, the quest for the conditions of responsibility is best understood as finding an answer to this latter question. The hidden idea, which makes the normative interpretation appealing is that certain forms of holding people responsible go beyond holding certain judgments: they involve a demand for justification, apology or compensation or they constitute a form of sanction or punishment, the fairness of which has to be guaranteed. But, as we will see soon, one might resist this characterization of blame.

Again, the present problem of attributionism is that the Control Principle apparently renders the theory unfair. To rebut this claim, attributionists have to come up with a concept of moral appraisal which can be fairly applied even if the agent did not exercise control over the thing assessed. To do so, attributionists develop a rather weak notion of moral criticism, by means of which they can avoid at least certain kinds of objections based on the fairness of responsibility-attribution. But before turning to their characterization, first I will present what may be the weakest notion of moral criticism which does not yet deny altogether the possibility of appropriately evaluating other's moral faults. After then I will confront it with attributionist accounts of blame.

8.1 Living without Responsibility

In "Determinism *al Dente*", Derk Pereboom argues from a hard incompatibilist standpoint that the truth of determinism (just as quantum-indeterminism) undermines moral responsibility. However, Pereboom argues that, contrary to what libertarian and compatibilist authors tend to claim, living without free will and responsibility is not that terrible, since "although we therefore never deserve blame for having performed a wrongful act, most moral principles and values are not thereby undermined" (p. 22). In reviewing the remains of morality in a deterministic world, Pereboom pays special attention to blame, since this practice seems to be the most evidently vulnerable to the absence of responsibility:

> A very prominent feature of our ordinary conception of morality that would be undermined if hard determinism were true is our belief that people deserve credit and praise when they deliberately perform morally exemplary actions, and that they deserve blame when they deliberately perform wrongful actions. To deserve blame is to be morally liable to blame by deliberately choosing to do the wrong thing. Hard determinism rules out one's ever deserving blame for deliberately choosing to act wrongly, for such choices are always produced by processes that are beyond one's control. (1995, p. 32)

It is worth noting that what Pereboom refutes is a desert-based model of holding responsible, which presupposes control: "for an agent to be *morally responsible for an action* is for this action to belong to the agent in such a way that she would deserve blame if the action were morally wrong, and she would deserve credit or perhaps praise if it were morally exemplary" (Pereboom 2001, p. xx) According to his theory, since as a consequence of determinism we lack the relevant kind of control over our decisions, our ordinary practice of blaming is theoretically irrational and morally wrong (Pereboom 1995, p. 33). What are we left with then? Should we treat people as “earthquakes and epidemics” (ibid.)? Pereboom is just a bit more permissive:

> One still might legitimately have a feeling of moral concern about what persons do, or about what persons who are reasons-responsive do, which would differ from one's attitudes to earthquakes and epidemics. This feeling would be legitimate supposing it has no cognitive component that conflicts with hard determinism. (Pereboom 1995, pp. 33–34)

The underlying idea of the passage seems to be that although there is no substantive difference between agents and earthquakes regarding their freedom or responsibility, given the special abilities human beings possess and the significance they play in one another’s life, it is justified to have special concerns and feelings regarding their actions. However, this is not all we have according to Pereboom. Perhaps it is not that embarrassingly little what we can say about an earthquake. Since our moral evaluations of right and wrong remain perfectly meaningful in a deterministic universe, these judgments can also be meaningfully articulated and communicated. Moreover, as opposed to earthquakes, these criticisms might also be formative:

> Instead of blaming people, the determinist might appeal to the practice of moral admonishment and encouragement. One might, for example, explain to an offender that what he did was wrong, and then encourage him to refrain from performing similar actions in the future. One need not, in addition, blame him for what he has done. The hard determinist can maintain that by admonishing and encouraging a wrongdoer one might communicate a sense of what is right, and a respect for persons, and that these attitudes can lead to salutary change. Hence, one need not hold the wrongdoer morally responsible for what he has done, but rather consider him responsive to moral admonishment and encouragement. (Pereboom 1995, p. 33)

8.2 How to Blame as an Attributionist?

I suspect most would agree that this type of moral evaluation cannot be called blame—indeed, it is a rather superficial kind of moral assessment. Pereboom would readily admit that: his point was to show what is left *without* blame—his account is consciously revisionary. In this respect (and only this is respect) Pereboom's account resembles to J. J. C. Smart's theory (1961), one of the most often cited consequentialist accounts of moral responsibility (for other versions see e.g. Brandt 1969 and Schlick 1939). Smart offers an admittedly revisionist account of moral responsibility and blame: according to him, only a purely consequentialist account of moral responsibility can make responsibility compatible with the truth of determinism. Consequently, all the non-consequentialist considerations involved in blaming practices (such as issues of freedom, desert and retribution) are necessarily untenable and should be given up in favor of the view that moral appraisal—"praise" and "dispraise", as he calls it—serves exclusively forward-looking, formative aims.

The charge of superficiality, however, is sometimes raised against theories which claim to capture our ordinary, non-revisionist understanding of blame and moral appraisal. A similar suspicion was expressed by Susan Wolf in her discussion of so-called "real self views", according to which "an agent is responsible only for those actions which are attributable to her real self, understanding an action to be attributable to one's real self only if in performing it one is at liberty to govern one's actions on the basis of one's valuational system" (Wolf 1990, p. 34). Wolf notably argued that real self views, represented primarily by Gary Watson's value theory, capture only a superficial sense of responsibility:

> When we say that an individual is responsible for an event in the superficial sense, we identify the individual as playing a causal role that, relative to the interests and expectations provided by the context, is of special importance to the explanation of that event. And when we praise or blame an individual in the superficial sense, we

> acknowledge that the individual has good or bad qualities, or has performed good or bad acts. But when we hold an individual morally responsible for some event, we are doing more than identifying her particularly crucial role in the causal series that brings about the event in question. We are regarding her as a fit subject for credit or discredit on the basis of the role she plays. When, in this context, we consider an individual worthy of blame or of praise, we are not merely judging the moral quality of the event with which the individual is so intimately associated; we are judging the moral quality of the individual herself in some more focused, noninstrumental, and seemingly more serious way. (Wolf 1990, pp. 40–41)

Although Wolf's criticism preceded both Scanlon's and Smith's accounts, her point, if valid, would obviously constitute a challenge also for their theories. Moreover, as I have previously mentioned, the term "responsibility as attributability" was first used by Gary Watson in "Two Faces of Responsibility", where Watson's aims to defend his theory from Wolf's objections. He puts forward the following answer:

> The significant relation between behavior and the "real self" is not (just) causal but *executive* and *expressive*. When thought or behavior are exercises of what Dewey calls an agent's moral capacity, they and their results are open to distinctive kinds of evaluation. These evaluations are inescapably evaluations of the agent because the conduct in question expresses the agent's own evaluative commitments, her adoption of some ends among others. To adopt some ends among others is to declare what one stands for. (Watson 1996, p. 270)

Watson is certainly right in emphasizing that *responsibility as attributability* is more than the mere assertion of a causal connection. Being morally responsible in the sense discussed implies that the agent's particular behavior give some kind of extra information which entitles others to draw further conclusions with regard to the moral qualities of the agent. However, what Watson tells about the nature of responsibility-ascriptions in the attributionist sense seems to confirm Wolf's diagnosis. Since, according to Watson, issues of punishment and sanction, moreover, the adoption of reactive attitudes belong to the territory of accountability—as independent of attributability—, moral criticism in the attributionist sense (which he also calls *areatic appraisal*) seems to be no more than an evaluative judgment concerning the moral quality of the agent.

One would think that attributionists will fight side by side to rebut Wolf's criticism. But this is far from being true. On the very contrary, Angela Smith, in her discussion of the debate, sides with Wolf:

> [On one interpretation of Watson's theory—R.A.] aretaic appraisals would report how a person stands with regard to certain standards of human excellence, but would not bring with them any further implications of fault or discredit, and would not support any demands for reasonable regard on the part of others. But if aretaic evaluations are simply "grading" evaluations, which do not imply any normative failure on the part of the person appraised, then Watson's insistence that they are nevertheless "deep" forms of assessment becomes considerably more dubious. (…)
> If this is the only kind of appraisal we can make of persons on self-disclosure or attributability views, then I think Wolf is right to complain that such views can account at best for a "superficial" kind of responsibility. (Smith 2008, pp. 377–378)

Smith concludes that after further examination areatic appraisals characterized by Watson turn out no "deeper" than J. J. C. Smart's notions of praise and dispraise. As I pointed out before, this poses no problem for Smart himself, since his aim is not to *describe*, but to *reform* our responsibility-attributing practices. And it is not particularly worrisome for Watson either, since the central thesis of his paper is that attributability is only one form of moral assessment, which should be distinguished from *accountability*, that is, issues of fair sanctioning and punishment. Wolf's criticism is valid, but misses the point: moral appraisals in the attributionist sense do not and need not have the kind of depth which she talks about.

But this line of thought will obviously be insufficient for Smith, who argues that responsibility of attributability can capture all the ordinary intuitions about responsibility and thus aims to provide a unified, non-revisionist account. Among Scanlon, Smith and Watson, Smith's project is far the most ambitious: she claims that her theory can accommodate all common sense considerations about responsibility, thus providing a unified account. Now it is time to explore it.

To develop her characterization of moral appraisal, Smith starts out by noting that moral criticism always concerns some rational activity on the agent's part. This observation helps to explain how moral criticism can involve the sort of "reactive entitlement" crucial to explain how it differs from mere negative of positive grading (the one we became familiar with in Pereboom's incompatibilist account and Smart's notion of praise and dispraise). Smith identifies the reactive entitlement in question with the demand of justification implicit in moral appraisals: "Moral criticism, by its very nature, seems to address a *demand* to its target. It calls upon the agent to explain or justify her rational activity in some area, and to acknowledge fault if such a justification cannot be provided." (Smith 2008, p. 381) That moral appraisal is distinctive in this sense can also explain, according to Smith, why we are morally responsible not only for our voluntarily chosen actions, but for all our judgment-sensitive actions and attitudes: since these are the products of the agent's rational activity, it is always sensible to call upon her to explain and justify them.

Also, reactive entitlement is what gives moral criticism the special "depth" which Wolf claimed to be missing from real self views. However, as Smith rightly recognizes it, rational activity extends far beyond the moral domain. We can just as sensibly and appropriately address a demand of justification toward someone on the basis of her prudential, political or aesthetic judgments. Smith is ready to embrace this somewhat peculiar consequence of her account:

> I am just as *morally responsible*, and just as deeply morally responsible, for an imprudent action as I am for an evil one. What makes me morally responsible in both cases is that these actions reflect my assessment of reasons, and therefore I can, in principle, be called upon to defend them and am open to rational (and in some cases moral) criticism if an adequate defense cannot be provided. (Smith 2005, p. 385)

But then, consequently, to distinguish moral criticism from other forms of rational criticism we have to introduce another factor. Smith argues that moral criticism has a special *significance* because of its *content*: since moral demands define, in a somewhat Scanlonian spirit, what we owe to each other, the violation of these demands has a direct bearing on those who have been wronged. While other kinds of rational criticism leave the personal relationships of the criticized intact, moral criticism expresses impairment in the quality of these relationships. This is why reactive sentiments are warranted: resentment and indignation are reasonable responses, because moral criticism tracks moral violations which leave someone being wronged.

These thoughts about the depth, content and significance of moral criticism recurs elsewhere, and seem to be the consensual view of attributionist authors and those inspired by their thoughts. That blame gains its special force and can be distinguished from other kinds of criticisms by its descriptive content was first claimed by Pamela Hieronymi in her article “The Force and Fairness of Blame” (2004). Hieronymi attacks a claim akin to the Control Principle, which she calls the “target charge of unfairness”, i.e.,

> that blaming a wrongdoer can be unfair because blame has a characteristic force, a force which is not fairly imposed upon the wrongdoer unless certain conditions are met—unless, e.g., the wrongdoer could have done otherwise, or is able to control

> her behavior by the light of moral reasons, or played a certain role in becoming the kind of person she is. (Hieronymi 2004, p. 115)

Since Hieronymi does not intend to specify the conditions which, according to the target charge of unfairness, render blame unfair, it seems to be a natural suggestion to say that the Control Principle is one version of the claim Hieronymi objects to. She argues, following Peter Strawson, that blame essentially involves the judgment that the agent displayed ill will or disregard and that this judgment in itself can account for much of the special force of blame. The significance we attach to blame and its subsequent force derives from our expectation to be in relationships of mutual regard and from the consequences, which moral failures bring into these relationships:

> It seems quite plausible to me that standing in relations of mutual regard is of considerable importance to creatures like us. Thus the content of a judgment of ill will can carry a certain amount of force—despite being descriptive. (…) A change in what you or another person thinks about the quality of your will, in itself, changes your relations with them. (…) That judgment—even if incorrect—makes it the case that you no longer stand in relations in which your good will is recognized on all sides. Thus the force of a judgment of ill will, I suggest, derives from the importance of standing in relations of recognition of mutual regard. The force of a judgment of ill will is found in and carried over from its content, even if the content is merely descriptive. (Hieronymi 2004, p. 124)

Hieronymi argues that the force of blame, which reactive attitudes carry, should be traced back not to their affective component, but to the complex set of judgments constitutive to these emotions. So, at the end of the day, although blame involves more than judgments (reactive attitudes are obviously affective), the *force* of blame is to be found in a certain set of descriptive judgments.

Finally, Thomas Scanlon's account of moral appraisal and the subsequent conception of responsibility is almost identical to Smith's theory. First, Scanlon agrees that being morally responsible “in the most general sense” means that “we can in principle be called on to defend these attitudes with reasons and to modify them if an appropriate defense cannot be provided” (Scanlon 1998, p. 272), and that, accordingly, we are responsible for our judgment-sensitive attitudes. Scanlon, just as Smith and Hieronymi, admits that moral appraisal makes certain emotional responses appropriate, and he insists, just like the other two authors, that the weight of blame is to be located in the significance of its descriptive content:

> To see the special force of the kind of self-reproach that guilt in the narrow sense involves, on my view, consider first the significance, for other people, of the moral criticism on which this reproach is based. If an action is blameworthy, then the agent has either failed to take account of or knowingly acted contrary to a reason that should, according to any principles that no one could reasonably reject, have counted against his action. So, in addition to whatever loss this action may have caused, the agent’s mode of self-governance has ignored or flouted requirements flowing from another person’s standing as someone to whom justification is owed. (Scanlon 1998, p. 271)

It might be worthwhile to note that recently Scanlon (2009) offered a somewhat different characterization, which, however, partly retained the emphasis on the judgment-like structure of blame (although here it is supplemented with a responsive-attitudinal component):

> Briefly put, my proposal is this: to claim that a person is blameworthy for an action is to claim that the action shows something about the agent’s attitudes toward others that impairs the relations that others can have with him or her. To blame a person is to judge him or her to be blameworthy and to take your relationship with him or her to be modified in a way that this judgment of impaired relations holds to be appropriate. (Scanlon 2009, pp. 128–129)

Now that we have seen the particular accounts of moral appraisal and blame provided by attributionists, we are in a better position to summarize the criteria such accounts have to meet. Most generally speaking, the central task of attributionists is to provide a characterization of moral criticism which is neither *too weak*, nor *too strong*.

An account of moral criticism can be too weak in two ways. On the one hand, as we have seen in Pereboom's case, if we understand moral criticism as simply a moral evaluation of certain events, character traits, decisions, etc., then it is unclear how these evaluations differ from the ones we commonly apply to things and natural events (cf. earthquakes). But even if we accept Smart's consequentialist proposal, it will still make sense to "dispraise" children and non-human animals as long as holding them responsible in this sense has beneficial future consequences. This is one of the main reasons why consequentialist accounts of responsibility are not quite popular: any conception of blame should account for the fact that only adult human beings with normal capacities (in a yet unspecified sense) are proper objects of it.

On the other hand, and this is a more pressing issue in the present context, moral criticism has to be distinguished from other kinds of rational criticism. Since attributionists maintain that moral criticism differs from "mere grading" because it targets the agent's rational activity, which is called for justification or revision by the criticism, they have to explain how *moral* criticism is different from addressing a demand to justify or revise one's prudential, aesthetic or philosophical views. The common attributionist answer emphasizes the *significance* of moral norms and standards in each other's life and argue that moral criticism, as opposed to other forms of rational criticism, implies an impairment in the personal relationships of the blamed person.

But the conception of moral criticism can also be too strong. Here I would like to return to the distinction which I have previously made between the descriptive and normative adequacy of a given concept of blame or moral criticism. With regard to descriptive adequacy blame can be too strong if it presupposes the obtaining of such facts which, given our present world, cannot obtain. Although attributionists are usually not particularly interested in the metaphysics of the free will debate (and this might be an understatement), both Scanlon and Hieronymi point it out that according to their conception of blame the truth of its constitutive judgments does not depend on the truth or falsity of causal determinism.

And finally, blame or moral criticism can be too strong, if it is not a morally justified response to facts about being responsible. Naturally, whether a concept is too strong in the normative sense depends on how authors define the scope of responsibility and what normative considerations they think have bearing on the moral justifiability of applying this conception. So we returned to the schema of normative interpretation and to the question how proponents of attributionism fill in its three variables (extension, fairness, responsibility-attributing practices). So far we have seen that one of the main appeals of attributionism is that it has a particularly wide scope, which makes it able to explain and justify problematic cases of responsibility (attitudes, negligence). In this chapter we encountered the judgment-based conception of moral criticism embraced by attributionists, which locates the special force or depth of moral criticism partly in the implicit demand expressed by it, and partly in the significance which morality plays in our life. Now we are ready to examine whether attributionists are right in claiming that we can fairly criticize people for things beyond their control.

8.3 The Fairness of Moral Appraisal

The charge that moral criticism might be unfair when applied to problematic cases is recognized by all the attributionist authors mentioned so far. In the previous chapter we already encountered Angela Smith's answer, i.e., that from the moral inappropriateness of certain forms of holding responsible we cannot infer that the agent is not responsible. Also, now it might be easier to see why this objection fails to undermine the normative interpretation: although Smith’s characterization excludes many forms of what we ordinarily mean by holding someone responsible, e.g., sanctions and punishment, her conception of being responsible is still defined by a responsibility-attributing practice, i.e., moral appraisal, which addresses a demand to the agent to explain, justify or revise her evaluative judgments.

Scanlon and Hieronymi use a different strategy as far as they accept that the charge of unfairness is a legitimate one, but argue that it cannot be applied to the discussed concept of blame. The central move these defenses take is to recognize that blame is constituted by nothing but a set of judgments. Moreover, blame can remain unexpressed: according to these authors simply holding certain judgments is enough to blame someone. But if this is the case, then it is hard to see how blame could be unfair as long as the judgments it entails are true.

Hieronymi puts forward the following argument for the conclusion that true judgments cannot be unfair: If certain judgments could be unfair in themselves then we would have a *prima facie* epistemic obligation *not to hold* them. But even if we put aside the problem that we cannot change our judgments at will, this obligation seems incredible. Of course, there might be unfair ways *to come to know* a judgment—for instance, if we violate someone’s right to privacy on the way. And also, we might be unfair in *forming the judgment*—if we base our judgment on insufficient reasons or evidences. But these are not the forms of unfairness at issue: the exact charge is that blame is unfair *to the agent* (if certain conditions are not met), because it puts unjustified burdens on her. Hieronymi seems to be right to conclude that “the fairness of suffering such burdens will turn on the accuracy of the judgment of which they are an immediate consequence” (Hieronymi 2004, p. 131).

Scanlon distinguishes two versions of the claim that “insofar as people’s actions are due to causes outside them it is unfair to blame them for acting as they do, since they cannot avoid acting that way” (Scanlon 1998, p. 282). The first of them concerns accuracy: “It is unfair to condemn a person for a certain action if that condemnation is based on inaccurate or incomplete information, when a fuller or more accurate account would reveal that the person is not as bad as he is being portrayed" (Scanlon 1998, p. 283). This objection would apply, e.g., to agents with deprived childhood or victims of other, unfortunate formative circumstances. As Scanlon points it out, this objection gains its appeal from the fact that the agents in question seem to be innocent in becoming the (bad) person they are. However—so Scanlon argues—we have no reason to accept that in order to appropriately evaluate someone, we have to hold her responsible for having developed the character traits in question[23].

The second formulation of the fairness-claim holds that "it is unfair to blame a person for acting in a certain way if he or she has not had adequate opportunity to avoid this condemnation" (ibid, p. 285). Scanlon offers two independent answers to this objection. On the one hand, he argues that the objection is legitimate only if we presuppose that "the criticism in question inflicts some cost on the person judged" (ibid.). But as we have seen, Scanlon understands moral criticism as a judgment, and forming a judgment is arguably not the kind of thing which would take such burdens on the blamed one:

> In considering the conditions under which moral criticism is appropriate, what we are concerned with is the appropriateness of the judgment that a person has acted wrongly, not the appropriateness of engaging in any particular form of blaming behavior, such as admonishment, shunning, or the withdrawal of friendship. (Ibid.)

On the one hand, since judgments in themselves cannot be counted as sanctions, their justifiability or appropriateness depends on whether they are true or false. Thus, if we accept that these judgments of agents can be true—what can hardly be doubted—, there is no room

[23] Smith introduces the very same distinction: "I think we would do well to distinguish two different questions: the question of one's responsibility for becoming a certain sort of person, and the question of one's responsibility for the judgments expressed in one's actions and attitudes" (Smith 2008, p. 389).

for the charge of unfairness anymore. Moreover, on the other hand, Scanlon points it out that the last form of blaming behavior, i.e., "the withdrawal of friendship" seems not to raise issues of fairness at all:

> it does not generally seem unfair to react to a malefactor's actions in ways that adversely affect his interests. Imagine, for example, an incorrigible opportunist and liar who takes advantage of everyone he can, and suppose that these characteristics are due to his miserable childhood, in which everyone he encountered behaved in this way and he had little choice but to do anything he could to survive. It does not seem to follow that it would be unfair to avoid dealing with this person or entering into relations of friendship and trust with him. If this is not unfair, why should there be any objection on grounds of fairness to the judgment that he has acted wrongly? (Scanlon 1998, p. 285)[24]

As we have seen, proponents of the attributionist view follow more or less parallel arguments to rebut the charge that the scope of responsible agency which they propose is wider than the one for which we can fairly hold people responsible. That this poses a problem for both Scanlon and Hieronymi shows that both of them accept some version of the normative interpretation. And, although neither of them formulates the charge of unfairness in terms of control, we can confidently assume that their answers, if they are viable, can be applied also to this formulation. The solution they propose relies on a judgment-based analysis of blame and moral criticism, which is detached from overt forms of holding responsible such as expressing reactive attitudes, sanctioning or punishing. Since it is hard to see how holding

[24] Pamela Hieronymi puts forward the very same claim:

> [I]t might seem similarly unfair that someone who has become a generally unreliable person, due to formative circumstances outside her control, should be systematically subject to the burdens of being constantly distrusted. Yet it is not clear that those who interact with her can be charged with unfairness in distrusting her. In fact, if a person will always let you down, that seems to be solid grounds for distrust rather than a condition under which distrust would be unfair. (Hieronymi 2004, p. 119)

true judgments could be unfair, Scanlon and Hieronymi go on to argue that the charge of unfairness cannot be applied to their accounts of responsibility.

8.4 Objections

One can contest on at least three grounds the sustainability and relevance of the discussed concepts of moral appraisal and blame. First, it can be argued that the judgment-based model of blame misconstrues the phenomenon. Most notably, some authors (e.g. Mason 2011, Clark n. p.) argue, following Peter Strawson, that affective responses—more specifically reactive attitudes—are necessary constituents of blame. As we have seen, attributionists admit to the somewhat vague idea that moral appraisal makes certain emotions *appropriate*, but they deny that these emotions would be necessary to blame someone.

Second, one can say that although these characterizations of blame make perfect sense, they capture only a marginal phenomenon. This insight is expressed by Randolph Clarke, who writes:

> The moral appraisals at issue in this central area concern moral demands or requirements, noncompliance with which is one thing to which the overt treatment responds. Critical assessment of this sort is blame only if it includes something more than cognitive judgment, some kind of disapprobation. (…) Somewhat more peripherally, we appraise each other with respect to our non-voluntary responses to obligation-related reasons. And more peripherally still, we issue judgments concerning whether thought or conduct that expresses an agent's evaluative commitments meets ethical standards other than obligations. (Clarke, n. p.)

Note that this criticism has more severe consequences for Smith than for Watson or Scanlon. Whereas the latter admit that responsibility as attributability is only one face of responsibility and thus it cannot capture all the important characteristics of our practices, Smith's project is to develop a non-revisionary and unified concept of responsibility. If Clarke is right, then this attempt is doomed for failure.

Finally, it can be questioned that moral appraisal can be detached from overt practices of holding responsible. Although he uses a somewhat different terminology, Robert Adams can be interpreted as expressing this suspicion:

> To me it seems strange to say that I do not blame someone though I think poorly of him, believing that his motives are thoroughly selfish. Intuitively I should have said that thinking poorly of a person in this way is a form of unspoken blame. (Adams 1985, p. 21)

Although Adams admits that involuntary sins might deserve different kind of treatment than voluntary ones, our reactions (such as reproach) to blameworthy thoughts and emotions cannot always be sharply distinguished from forms of punishment. To sum up Adams's point, judgmental forms of moral criticism cannot be conceived independently of our blaming practices. Our responses to morally objectionable actions and attitudes, from negative evaluation to the adoption of reactive attitudes and reproach to sanction and punishment, show a continuity which the concept of moral criticism discussed above seems to break. Clarke raises a somewhat similar concern, although he focuses on the justificatory relation between what he calls "appraisability" and overt responses to someone's faulty actions and attitudes:

> We have a view of each other as morally responsible agents and a practice of assessing and responding to those whom we take to be responsible for specific things. At the center of this practice, we form appraising attitudes *the propriety of which is inseparable from the justification of overt responses*; what grounds the one provides at least partial grounds for the other (Clarke n. p.)

This last form of criticism creates a more serious threat to attributionist accounts than the former two. On the one hand, it suggests that any attempt to distinguish different concepts or faces of responsibility is necessarily destined to failure: since the seemingly diverse forms of responsibility-attributing practices cannot be meaningfully separated, we need a uniform theory which is able to handle every practice within the moral sphere (arguably, legal punishment is still a separate question). But if this is so, then the attributionist strategy of sidestepping issues of fairness by providing a relatively weak, judgment-based account of moral criticism cannot be viable either: since many overt forms of holding responsible do raise considerations about fairness, attributionists have to give a substantive answer to the question how holding people responsible can be fair under the conditions they have defined.

Chapter 9: Fairness

Neither Smith, nor Scanlon says much about what they mean by fairness, presumably because they assume that their concepts of moral criticism and blame are immune to the charge of unfairness under any interpretation. Nevertheless, to make the current project complete, I need to briefly discuss what concepts of fairness are in play when saying that it is unfair to hold people responsible if they did not exercise voluntary control. In the following I will rely heavily on R. Jay Wallace's ideas on the issue, although I will supplement them. I do not wish to give a detailed elaboration of the normative concepts in question: my goal is simply to identify certain moral intuitions underlying the fairness claim. Although the intuitions I will discuss and the subsequent notions of fairness seem to be independent of each other, after a careful exploration of the concepts and taking into account considerations external to the present context, it might turn out that they collapse back into one or two separate concepts.

Distinguishing different concepts of fairness can also illuminate the separate and distinctive ways by means of which the lack of control excuses or exempts the agents from responsibility. Also, I will discuss Scanlon's remarks about what he calls the Desert Thesis to illuminate some further consequences of the normative interpretation, which will be the subject of the next chapter.

9.1 Incompatibilists: Fairness as Desert

Although my discussion focuses mainly on compatibilist theories, at this point it is worthwhile to briefly return to the original debate between compatibilists and incompatibilists and examine the concept of fairness in this context. If the normative interpretation is correct, as I claim it to be, then these positions should also be reformulated. The incompatibilist's claim, i.e., that causal determinism excludes free and responsible agency should be understood as follows: if causal determinism is true, it is always unfair to hold people responsible. But how does exactly the incompatibilist argues for this claim?

Saul Smilansky (2008) is one of the very few authors who attempted to characterize the incompatibilist position in terms of fairness. His main contention is to highlight—consistently with his general position, which he labels as Fundamental Dualism (p. 140, see also Smilansky 2002)—how compatibilist and incompatibilist intuitions can both independently arise in certain situations. To demonstrate this he puts forward a thought experiment, which provokes robust compatibilist-friendly intuitions, while leaving intact the most basic incompatibilist worries. In his proviso four people, A, B, C and D are stuck on an island where they have to take a long, effortful and dangerous road everyday to reach their food supplies. On the first day they take the road together. But, the story continues,

> On the following day, D refuses to come with A, B and C: the journey is hard, and there is some danger; he prefers to stay on the beach. They threaten that they will not give him any on the supplies they bring back, but he does not believe that they will not share things with him. And indeed, when they return, they cannot resist his pleas of hunger, and share sprains his ankle just as they return to their camp, and cannot continue to make the daily journey. A and B are hence left to do all the work by themselves. Each night, when they return, they share the provisions they carried with C equally. To D, they give only water and the minimum amount of food that will keep him alive. (Smilansky 2008, pp. 239−240)

Is it fair to treat D this way? Smilansky’s answer is: yes and no. On the one hand, “civilized human existence” (p. 241) requires us to make other people’s treatment at least partly dependent on such notions as choice and control, understood in a compatibilist manner. On the other hand, however, this cannot blur our initial worries which pulled us toward incompatibilism. In a situation like this the incompatibilist’s main contention is this: Is it fair to treat D differently from C? Is it fair to sanction him with hunger and imprisonment?

Standard incompatibilist arguments, according to my understanding, have the following structure:

> P1: Holding someone responsible for doing *x* is fair only if condition *C* applies.

(Where *C* is something like “she had adequate opportunity to avoid doing *x*”, “she possessed the ability to refrain from doing *x*”, “did *x* freely”, “exercised control over *x*’s coming about”, etc.)

> P2: If causal determinism is true, *C* never applies.
>
> P3: Causal determinism is true.
>
> P4 (from P2 and P3): *C* never applies.
>
> C (from P1 and P4): It is always unfair to hold someone responsible for doing *x*.

To determine the exact content of P1, we obviously has to return to the question which was raised in the previous chapter: how should we characterize the constitutive practice of holding responsible? Most incompatibilist authors assume that the concept of moral responsibility should be able to provide justification to certain retributive practices, in particular to certain forms of punishment. This does not mean that the agent’s responsibility should be the *only* factor determining the fairness of punishment—most incompatibilist would admit that forward-looking considerations play a significant role here. These considerations, however, according to the incompatibilists, cannot override what retributive justice requires from us—there are responsibility-based constraints on the fairness of punishment. Consequently, one cannot simply put responsibility-related considerations aside and substitute them with, let’s say, a consequentialist concept of fairness. As Smilansky puts it,

> It might be argued that nothing much would happen were we to become hard determinists, for we could then mold our social environment along utilitarian or contractual lines, and surely those would enable us to track considerations of fairness (or their equivalents). [But] under hard determinism, it is not at all clear that *any* such practices distinguishing among people can be justified. (…) By jumping at this point aboard some other and very different (e.g., utilitarian) train, a hard determinist would be betraying the insights of his position. (Smilansky 2008, p. 242)[25]

Obviously, these are only the contours of the incompatibilist's argument—how convincing we will find it mostly depends on the presently ignored details. What most incompatibilist authors seem to agree on is that there is a basic, retributive aspect of responsibility-attributing practices, captured by the notion of *desert*, which is independent of consequentialist or contractual considerations and which, under causal determinism, renders holding responsible universally unfair. No one deserves to be punished, incompatibilists say.

Compatibilists might use several strategies to challenge the incompatibilists' line of argument. The most direct way is to refute either P1 or P2. One way to do so is to demonstrate that the *C* condition to which the incompatibilist is committed, based on our intuitive judgments of responsibility, is not a necessary condition of moral responsibility (P1 is false)—thus nothing

[25] A similar objection is raised by Lene Bomann-Larsen (2010):

> Strong revisionists object that consequentialism is unfair because it, too, *ex hypothesi* entails punishing the innocent. Now, consequentialists may admit this, while maintaining that, unlike retributivists, *they* do not go about punishing the innocent while *pretending* that she is guilty, they merely say that out of concern for the common good we should minimise the occurrence of harmful actions, and maintain the institution of punishment in order to achieve that goal. (…) But this is unsatisfactory. Punishment on this score is rather a form of 'telishment'. And as Pereboom notes, to demand punishment of some (by definition) innocent individual in order to maximise utility amounts to harming people as a means to collective well-being. (Bomann-Larsen 2010, p. 5)

stands or falls on whether it is compatible with the truth of determinism. Alternatively, the compatibilist can admit that *C* is indeed a necessary condition of responsibility, but go on to argue that condition *C*, correctly understood, *is* consistent with the truth of determinism – thus P2 is false. I take it that the often frustrating and never ending dispute between classical compatibilists and incompatibilists about the proper analysis of "cans" and abilities exemplifies this latter strategy.

Another method to neutralize the incompatibilist's fairness objection is to argue against retributivism. This is the line of argument which Scanlon follows:

> It is sometimes said that feeling guilty for having done something necessarily involves the belief that one should be made to suffer in some way for having done it. (…) Let me call the moral idea underlying such claims—the idea that when a person has done something that is morally wrong it is morally better that he or she should suffer some loss in consequence—the Desert Thesis. (…) If the "ordinary" notions of guilt, blame, and so on do indeed have this desert-entailing character, then the account of moral criticism that I am defending is also in this respect revisionist. The reasons for my revisionism (if my view is indeed revisionist) have nothing to do with concerns about free will. To my mind, no degree of freedom or self-determination could make the Desert Thesis morally acceptable. (Scanlon 1998, pp. 274–275.)

What I find especially interesting about Scanlon's proposal is that, similarly to Smilansky's previously cited remarks about utilitarianism, it comes exclusively from the normative realm. Scanlon does not want to deny that in certain respects his account of moral criticism (and consequently that of moral responsibility) is revisionist – as the quote suggests, he is willing to admit that we ordinarily attribute "desert-entailing character" to many of the related concepts. He finds the Desert Thesis *morally* unacceptable – and this is sufficient reason for him to refute any such concept of responsibility which presupposes the thesis.

Scanlon's insight highlights a further consequence of the normative approach which I defended in length in Chapter 7. Determining the conditions of moral responsibility is an essentially normative task in a further sense: we cannot accomplish this task without committing ourselves to certain normative principles and ideals. If being morally responsible is being the fair target of certain responsibility-attributing practices (as the normative interpretation claims), then in order to complete our theory we need to determine which moral principles regulate the fairness of these practices. At the end of the day our theory of responsibility cannot stay completely neutral and noncommittal with regard to normative ethical questions.

I find this last conclusion significant and it's worth keeping it in mind when assessing the pros and cons of particular accounts of moral responsibility. Among other things it follows from it that there might be substantial disagreements between theories of moral responsibility with an exclusively normative ethical basis. Let's return to Scanlon! He refuses the Desert Thesis. If someone with retributivist sympathies comes to him and asks: "but how can punishment be justified if no one ever deserves to be punished?", predictably he will come up with a contractualist justification of punishment and try to talk the other out of his retributivist views. At this point they are in the middle of a conversation about substantial ethical issues which, at least according to the traditional view of these matters, has little to do with the problem of free will and moral responsibility.[26]

I will explore only one counterclaim retributivists might make. They can argue that we need not accept the Desert Thesis in order to make sense of the incompatibilist's complaint—it is enough embrace the rarely debated claim that it would be unfair to punish the innocent. It is

[26] That theories of moral responsibility are not neutral with regard to normative ethical theories might become obvious, if we recall that J. C. C. Smart's consequentialist account of responsibility, as critics have rightly pointed out, ran into the same old problems as utilitarian moral theories usually do.

controversial whether this would be sufficient to define an independent standpoint. Bomann-Larsen, for instance, writes,

> For retributivism to be a *distinct* justificatory position it must involve something more than the negative claim that it is morally wrong to punish the innocent; it must involve the positive claim that it is *also* morally right to punish the guilty, that it is somehow owed to *them*. Though the stronger claim includes the weaker claim, the weaker claim is not particular to retributivism, but can be seen as a general constraint on any form of punishment: punishing the innocent is the paradigmatic instance of injustice. (Bomann-Larsen 2010, p. 4)

But even if accept that this is all the retributivist has to say, why would we also concede that, given the truth of determinism, everyone is *innocent*? Is this further step needed in order to make the incompatibilist's argument complete?

It seems so. So far the incompatibilist maintained that it is unfair (because undeserved) to hold someone responsible unless certain conditions (which I called condition *C*) are met and that these conditions cannot be met if causal determinism is true. But what makes certain responsibility-attributing practices, most importantly punishment, undeserved in the absence of these conditions? The most plausible answer is to say that acting in accordance with deterministic physical laws leaves everyone, in some sense, innocent, regardless of what and why they did.

This is a powerful idea for the first sight, but not one particularly easy to argue for. In *Responsibility and the Moral Sentiments* R. J. Wallace argues in length against what he calls the "generalization strategy", i.e., that "some (…) principle of fairness, required to account for concrete judgments of excuse in which we have great confidence, would equally tell against holding people morally responsible if determinism should be true" (Wallace 1998, p. 115). He identifies two notions of fairness, which might serve the aims of the defender of the generalization strategy.

The first concept which Wallace discusses is fairness as *desert*. As he points it out, the notion of desert concerns particular actions and omissions: we do not deserve to be held responsible not in general, but for particular instances of our agency. Given this localized nature of claims about desert Wallace goes on to argue that claiming that someone does not deserve something functions as an excuse, which, according to the incompatibilists, should be extended to every action and omission, if causal determinism is true. Then he develops an account of excuses in the spirit of J. L. Austin (1956), according to which excuses suspend our judgments of responsibility, because they show that the agent, after all, hasn't done anything *wrong*, i.e., she hasn't violated her moral obligations. Finally he presents a moral principle, which he calls the "no blameworthiness without fault" principle, according to which "people do not deserve to be blamed if they have not done anything wrong" (Wallace 1995, p. 135). This is the principle, Wallace argues, which underlies our judgment that a valid excuse makes responsibility-attribution unfair.

However, we have no reason to think that causal determinism would make any difference in this respect. Wallace provides an account of obligations, according to which one can violate her obligations by manifesting a bad quality of will or choice[27]. Excuses apply when, contrary to appearance, one did not manifest such objectionable attitudes and consequently did not violate her obligations. But, as a reasonable incompatibilist would instantly admit, the truth of determinism has no bearing whatsoever on people's possessing bad attitudes and their being manifested in actions and omissions. The generalization strategy—supplemented with the "no

[27] Again, this is a distinctively Strawsonian feature of Wallace's account.

blameworthiness without fault" principle—thus fails, since people keep violating their obligations also in a deterministic world.

Wallace's account of excuses and its relation to desert is extremely subtle and here I gave only a sketchy overview of its structure. However, some important advantages are already apparent. First, the "blameworthiness without fault" principle, as it stands, seems basic and uncontroversial. And second, it really seems to capture our intuitions about what makes responsibility-attribution unfair. Very roughly, holding people responsible, if certain conditions haven't been met, seems to be parallel to punishing the innocent. I find that this basic insight is adequately expressed by the principle.

However, this account also shows that accepting the normative interpretation requires different treatment of certain issues than how they are usually conceived. I think, almost anyone would agree with Wallace that it is undeserved to hold responsible someone if she is not blameworthy. But proponents of the metaphysical interpretation would give a different explanation of why these agents are not blameworthy. Blameworthiness is usually characterized as the conjunction of two facts: that the agent did something wrong and that she is responsible for what she did. Excuses, according to advocates of the metaphysical interpretation, provide ground to deny the latter, but not the former. They would claim that since the agent wasn't properly connected to her action, she is not responsible for it and so she is not blameworthy.

But this option is obviously not open to Wallace or anyone endorsing the normative interpretation: since being morally responsible for something *is* being a fair target of responsibility-attribution, we cannot explain why it is unfair to hold someone responsible by stating that she is not responsible—this would make the explanation circular. Again, this is not to say that Wallace cannot make sense of the idea that a strong enough connection between the agent and her action has to be in order to make blame fair: but he is forced to say that the connection in question has a bearing on the *wrongness* of the action. Since, according to Wallace, we can violate a moral obligation only by manifesting a bad quality of will or choice, without this condition being met the agent hasn't done anything wrong, and so she obviously does not deserve blame.

It is important to note that this solution requires a substantive account of what moral obligations are. Wallace's defense of compatibilism relies crucially on the claim that we can violate our obligations solely by exhibiting a bad quality of will. Moreover, Wallace does not discuss the delicate issue whether obligations *themselves* are compatible with the truth of causal determinism. It is relatively easy to understand why causal determinism might pose a threat to moral obligations. Although its exact content is much debated, the "ought implies can" principle is widely accepted among moral philosophers. However, if (as most incompatibilists contend) causal determinism deprives us of the ability to do otherwise, then every time when we do not fulfill our obligations, it will be true that we *cannot* fulfill them – but then, according to the principle, it is not true that we ought to fulfill them. But obviously, we want "ought" judgments to be true also when the agent has not done what she ought to have done. Thus, incompatibilists can argue, the universal lack of the ability to do otherwise undermines "ought" judgments (for a subtle defense of this view see Haji 1999)[28].

Here I will not further elaborate on this issue or follow this half-imaginary debate. However, it might be useful to sum it up. The core incompatibilist idea, which I have explored in this chapter, is that the truth of determinism would render holding people responsible (at least as far as punishment and sanctions are concerned) universally unfair, because they do not *deserve* these treatments. Holding people morally responsible, if causal determinism is true,

[28] Haji, I. 1999. Moral Anchors and Control. *Canadian Journal of Philosophy* 29, pp. 175–203.

would amount to punishing the innocent, incompatibilists say. Wallace, by contrast, argues that this would be true only if it would follow from determinism that no one ever violates her obligations. But, since obligations are violated by manifesting a bad quality of will, and this might be done independently of the truth of determinism, the incompatibilists charge of unfairness fails. The incompatibilist, in turn, might answer to this (although this is not the only way of getting off the hook) that the truth of determinism would undermine not only judgments of responsibility, but also "ought" judgments—if determinism obtains, then no one is obliged to do anything, since we are unable to do anything but what we actually do. It is undeserved to punish or sanction people because they are innocent; and they are innocent not simply because they haven't violated their obligations, but because they have no obligations at all.

9.2 Incompatibilists: Comparative Fairness

When I introduced Smilansky's scenario in the previous subsection, I said that incompatibilists would raise the following questions concerning it: is it fair to treat D differently from C? Is it fair to sanction him with hunger and imprisonment?

Now it is important to note that these are two separate questions. We might *prima facie* approve D's imprisonment in itself, but disapprove that he gets different treatment from C. Or, on the contrary, we can say that although it might be fair to treat C and D differently, these treatments cannot involve starving or imprisoning someone else. Up until now I discussed the second claim, i.e., that if determinism is true, then certain responsibility-attributing practices are always unfair. This is not equivalent to saying that everyone should get the *very same treatment*, since it might be that certain milder forms of blame or moral criticism (such as those we encountered in the Chapter 8) *are* morally appropriate in the case of D, but not in the case of C.

But there is a different line of incompatibilist thought, which is rarely distinguished from the first strategy which I outlined. This is articulated by the first claim, i.e., that *independently of the form of treatment* causal determinism renders it unfair to treat people *differently*. This idea often seems to be a logical, if not inevitable consequence of the truth of the second claim. That is, if causal determinism leaves everyone innocent and thus undeserving of sanction and punishment, then this is obviously true of everyone to the same degree and consequently we all should get the very same treatment. As I indicated in the previous paragraph, this is not necessarily the case; since standard incompatibilist arguments focus on the justification of one particular form of holding people responsible, they leave it open whether other, milder practices might be morally appropriate (for some reason which we did not yet consider).

More importantly, questions about *comparative fairness* might arise without making any assumption about the innocence of the agent. To see this, let us briefly return to the problem of resultant and circumstantial luck, proposed by Nagel. My example of resultant luck was the negligent driver, who, because he forgot to check the conditions of the brake, hits a child in the first case, while makes no harm in the second. Nagel's observation is that we would assess the driver differently in the respective cases, although, with respect to what was within his control, he behaved identically. However, if we consider these moral assessments *on their own*, then they do not seem to be neither unfair, nor unreasonable. It is not the case that our assessment of the "harmless" driver was too mild or the other too harsh. Rather, what we feel, *after comparing* the two cases, is that the driver would deserve *equal* treatment in both cases, since his behavior was identical in every, morally relevant respect. Moral luck cases, especially resultant and circumstantial luck are dilemmatic—or, echoing Nagel, paradoxical—exactly because they confront our intuitions about *desert* in particular cases with considerations about comparative fairness, arising when we contrast the two situations[29]. Since the two cases seem to be alike in every morally relevant aspect (since the agent exercised the same form and amount of control), we want them to be *treated* alike.

But how would this amount to a general incompatibilist argument? Even if we accept that resultant and circumstantial luck are existing moral phenomena, they are obviously not universally present. In certain cases we are unwilling to treat like cases alike, but why would we think that *all* cases are alike?

Nagel gives us a hint when in a famous passage he introduces “causal luck”:

> If one cannot be responsible for consequences of one's acts due to factors beyond one's control, or for antecedents of one's acts that are properties of temperament not subject to one's will, or for the circumstances that pose one's moral choices, then how can one be responsible even for the stripped-down acts of the will itself, if they are the product of antecedent circumstances outside of the will's control? (Nagel 1979, p. 35)

In a nutshell, Nagel’s overall argument goes like this: we all agree that nobody can be morally assessed on the basis of what is beyond her control. However, we seem to frequently violate this principle: for, instance, when we judge people for their temperament or the consequences of their actions. Moreover, if we totalize these effects of moral luck, it will turn out that we have no principled way to distinguish between controlled and uncontrolled features of agency, between actions and happenings, and thus the “responsible self seem to disappear, swallowed up by the order of mere events” (p. 36).

[29] A similar, comparative notion of fairness is recognized by Nelkin (2009). Although she does not elaborate on what she calls the “interpersonal” notion of fairness, I suspect that this notion is more akin to what we might label as “fairness as equal opportunities”, than the less robust concept of comparative fairness (or justice) which I have presented here.

Now it is probably easier to see how the incompatibilist argument might get along. First it seemed that the problem of comparative fairness arise only in certain, well defined situations. However, Nagel’s skeptical argument shows that in a way *all* situations are alike: since we are all completely and constantly out of control, there is no way to justify different moral treatment.

Whereas the desert-based argument relied on the intuition that, given the truth of determinism, we are all *innocent*, here the conclusion is that we are all *equal*, since we are all victims of causal luck. Consequently, whilst in the first case compatibilist replies focus on the notions of desert, obligation, innocence and guilt, in the latter case a possible defense would elaborate on the notion of comparative fairness and the principles underlying it. This is a significant difference, which a complex compatibilist strategy should take into account.

9.3 Compatibilists: Fairness as Reasonableness

Up to now we were discussing claims about the unfairness of holding responsible which came from the incompatiblist camp. There is, however, another notion which Wallace indentifies: fairness as reasonableness. Although the exploration of this notion also takes part of Wallace’s rebuttal of the generalization strategy, a brief exploration of the concept reveals that this notion of fairness favors the compatibilist camp. To illuminate the idea behind fairness as reasonableness Wallace gives the example of punishing children:

> Consider an example: a young child does something morally wrong--lies to her parents, say, about whether she has cleaned her room. There may well be good reason to scold or punish the child in this situation, but I take it we would think it unfair to hold the child fully responsible for her deed, in the way we would ordinarily hold morally responsible an adult who lies for personal advantage. (…) [I]t would be unfair roughly in the sense that it would be unreasonable to treat the child as fully accountable in the first place. (Wallace 1995, p. 108)

If claims about desert functioned as excuses, then claims about unfairness understood as unreasonableness are like exemptions: they call into question that our blame is directed to an *accountable agent* in the first place. Wallace identifies what he calls the A-conditions of responsibility (the conditions of being a responsible agent) as a set of cognitive and affective capacities, which makes the agent able to grasp and apply moral reasons and regulate her behavior in the light of them. He finally presents the following moral principle to support the claim that it is unfair to hold someone responsible if she does not meet these criteria: "it is unreasonable to demand that people do something—in a way that potentially exposes them to the harms of moral sanction—if they lack the general power to grasp and comply with the reasons that support the demand" (Wallace 1996, p. 161).

I find the notion of reasonableness particularly perplexing and difficult, but here I won't go on to explore it. What I find especially significant in Wallace's proposal is that it highlights how intimate the relation is between considerations that bear on the appropriateness of *addressing a demand* and considerations regulating the appropriateness of holding someone responsible if she *fails to meet the demand*. Correctly understood, what renders blame unfair, if directed to (temporarily or permanently) non-accountable agents is that it was unreasonable to expect that they would fulfill their obligations in the first place. Fairness as unreasonableness, as opposed to fairness as desert, concerns responsibility-attribution only derivatively: its primary application is to determine *what* we can expect from *whom*.

That the "what" and the "whom" are both in need of clarification suggests that Wallace omitted to talk about another question about responsibility, where considerations of reasonableness frequently arise. A moral demand can be unreasonable if directed to someone who lacks the relevant capacities to meet that demand. But a moral demand can be unreasonable also in *general* if, given how we conceive the capacities of normal human adults, *no one* can meet it. This latter problem, caught up by George Sher in "Kantian Fairness", concerns the scope of things for which we can be *in principle* reasonably held responsible. And it is of special significance to the present discussion, since attributionists might be interpreted as answering this question rather than the former two. Here are some quotations strengthening this understanding:

> In order for a creature to be responsible for an attitude, on the rational relations view, it must be the kind of state that is open, in principle, to revision or modification through that creature's own processes of rational reflection. (Smith 2005, p. 256)

> [W]e are responsible for all our judgment-sensitive attitudes: that is to say, we can in principle be called on to defend these attitudes with reasons and to modify them if an appropriate defense cannot be provided. (Scanlon 1998, p. 272)

However, if this is true then attributionists accounts answer to a question which seems to be conceptually prior to considerations about the desert and reasonableness (in the first sense) of holding responsible. Defining the scope of things which we can reasonably ask people to regulate does not settle the question whether someone is morally responsible for something or not. While a moral demand in itself can be perfectly reasonable, since normal human adults are usually able to meet it, it might be unreasonable to address it to this particular individual (because she, temporarily or permanently, lacks the relevant abilities) or undeserved to hold her responsible for not meeting it (because one or more excusing conditions obtain).

It could be said that the second sense of unreasonableness just discussed is not a different form of fairness but only a generalized version of the first one which Wallace introduced. In some sense this is right: whereas Wallace's notion is served to determine whether it is reasonable to address a demand to *someone*, given her general abilities, this second sense determines whether it would be reasonable to address a particular demand to *anyone*, given the general abilities of human beings. Answering affirmatively to the former question implies that we gave a positive answer also to the latter one.

But, if we subscribe to Wallace's general ideas about how excuses work, then claims about fairness as desert also presupposes that we take the moral demand in question to be reasonable. Following Wallace we affirmed that charges about desert rely on the content of our moral obligations: it is undeserved to hold someone responsible, if she did not violate a moral obligation. But if this is so, then the desert of responsibility-attribution crucially depends on whether we have a moral obligation to regulate the particular behavior. And it is not enough to point to certain current practices which seem to underline the existence of such obligations: these obligations also have to be morally justifiable, defensible, which, in turn, depends on the reasonableness of the obligations themselves. To put it more simply: for responsibility-attribution for something to ever be deserved, it is necessary to show, that it is reasonable to expect or demand the agent to behave in certain ways in the first place.

The distinction introduced here between fairness as desert and fairness as reasonableness (which itself can be understood in two separate ways) fits nicely to the previous discussion in Chapter 6 about the different ways in which we refer to the lack of control as an excusing or exempting condition. There my example was someone who has an obligation to record a television program from a distance. I distinguished three reasons which might exempt the agent from responsibility is she fails to fulfill her obligation: (i) she lacks the cognitive or executive capacities to accomplish the task (ii) the device does not function properly (iii) there is no device which would make it possible for her to accomplish the task. I find that the first case can be easily identified as a charge about unreasonableness in the first sense, while the third one as an application of unreasonableness in the second sense. While it is unreasonable to address a demand to someone who (for no fault of her own) is not able to meet it, it is generally unreasonable to address a demand which no one can meet.

The second case is somewhat more complicated, since it is harder to see why it would amount to the claim that it is undeserved to hold the agent responsible, if she lost control over the device. That here the fitting between the characterizations of fairness and the particular excusing condition is less obvious might be a sign of the diversity of desert claims in general. However, if we follow Wallace in saying that excusing conditions function by denying that the agent has violated her obligation, because she did not manifest certain mental qualities, then the second case fits into this description. Intuitively, the reason why loosing control in this second sense diminishes responsibility is that the agent's evaluative judgments, incentives, motives, choices or will (pick your favorite term), due to the improper functioning of the device could not be manifested in her omission. Exercising control over an external object requires, at the minimum, some sort of sensitivity to our intentions from the object's part. When we loose control over something, this sensitivity gets lost, through no fault of our own. Accordingly, a broken device prevents our bad will to be manifested and thus to violate our obligations.

This solution is far from being uncontroversial: it assumes that some kind of *mens rea*—to use the corresponding legal term—is a necessary condition for violating a moral obligation. Although this is a contentious claim, the above formulation has the advantage that it is compatible with any kind of mental state which one think to be crucial in establishing the agent's responsibility.

Consequently, we have found three ways to interpret the Control Principle:

(a) It is unreasonable to hold someone responsible if she lacked the relevant capacities to exercise control.

(b) No one deserves to be held responsible if, due to external obstacles, she was prevented from exercising control.

(c) It is unreasonable to address a demand (and consequently to hold people responsible for not meeting it) if it is in principle beyond agents' control to meet it.

Also, now we can see that these concepts of fairness are not completely independent from each other. That is, for holding responsible to be fair in the first or second sense it has to be fair in the third sense. As George Sher recognizes it, "blame and the attribution of responsibility involve the retrospective endorsement of the moral demand" (Sher 2005, p. 187). Consequently, establishing the reasonableness of the demand logically precedes considerations about standard excuses (b) and exemptions (a). Also, as our ongoing discussion suggests, it seems adequate to interpret attributionist accounts as being primarily concerned with the reasonableness of moral demands (c).

Chapter 10: Normative Ethics and Theories of Responsibility: The Division of Labor

I suspect that the last three chapters could have been confusing for some of those familiar with the usual context of the debate about free will and moral responsibility. Questions and debates concerning the adequacy of certain normative ethical theories have arisen more or less frequently. This might seem like an illegal entry intro issues that are independent of our present topic.

I made it clear in Chapter 7 that crossing the boundaries between theories of responsibility and substantial ethical theories concerning what is right and wrong, is a necessary consequence of the acceptance of the normative interpretation. If being responsible is being a fair target of certain responsibility-attributing practices, then determining the conditions of responsibility will first require us to define those considerations which regulate the fairness of these practices. In the previous chapter I presented what I take to be the most prevalent notions of fairness, which theories of responsibility make use of.

Still, the division of labor between theories of responsibility and ethical theories is a perplexing issue, which deserves more attention than what it currently gets. In this chapter I will dig deeper into this topic to make it clear how different standpoints on this issue can be articulated and where attributionist accounts (especially Thomas Scanlon's) stand. Probably much of what I intend to say will seem too basic or obvious for some. Sometimes I will simply repeat some points which I have raised earlier. However, I find that these issues are so rarely discussed explicitly that attempts of clarification are highly warranted.

In order to gain a deeper understanding of the issue itself, first it is useful to distinguish three questions: (i) what are the necessary and sufficient conditions of being morally responsible? (ii) how should we morally evaluate certain instances of responsible agency? and (iii) what kind of moral responses are justified when facing morally blameworthy or praiseworthy instances of responsible agency? Question (ii) is most often taken to be the one of normative ethics, independent of the meta-ethical debate over moral responsibility. In order to tell if someone is responsible for an action, omission or attitude it is obviously not necessary to determine the rightness or wrongness of the action, omission or attitude itself.

However, it is still an open question whether those providing an answer for (i) should also provide an answer for (iii). By putting the question this way is, we can easily recap the debate between proponents of the metaphysical and the normative interpretation which I discussed in Chapter 7. Defenders of the metaphysical interpretations contend that the two questions are separated: to say that someone is morally responsible for something is compatible with virtually any claim about the proper moral response. I argued against this view in length and maintained that facts about being responsible are conceptually dependent on claims about the moral appropriateness of certain responses—if we try to pull the two questions apart, it becomes truly mysterious what facts about responsibility are.

Note, however, that (iii) is consciously left ambiguous. On the one hand, no theory of responsibility aim to determine what the proper moral response should be to any particular action, omission or attitude. To put it bluntly, determining whether an action, omission or attitude should be praised of blamed requires us to know if it is right or wrong. But only normative ethical theories can answer this question—thus we returned to (ii). We can give a full answer to (iii) only by answering *both* (i) and (ii). This might seem a bit too abstract, but an example will make this point clear. Derk Pereboom, whom I have cited several times, defines being responsible the following way: “for an agent to be *morally responsible for an action* is for this action to belong to the agent in such a way that she would deserve blame if the action were morally wrong, and she would deserve credit or perhaps praise if it were morally exemplary" (Pereboom 2001, p. xx) That is, facts about responsibility constitute one of the necessary components which determine what the proper moral action should be in response to the given action (or omission or attitude): *if* conditions of responsibility are met *and* the action was wrong, *then* the agent deserves blame (and the same goes for praise). Note that this way of phrasing matters does not go against the spirit of the normative interpretation. It is still the case that facts about being responsible are defined in terms of the moral appropriateness of certain kinds of responses—what theories of moral responsibility do not want to specify is only whether these responses in the particular case should be negative, positive or morally neutral.

So far everything seems neat. But this is only because we consciously ignored a crucial fact: that the moral responses the appropriateness of which facts about responsibility determine are also actions and attitudes themselves, the wrongness or rightness of which, according to the picture just sketched, should be determined by our favorite normative ethical theory, not by any theory of responsibility. That is, whatever theory we choose to answer question (ii), it will also regulate the rightness or wrongness of responsibility-attribution itself. But this should be the task, according to the normative interpretation, of our theory of responsibility. Normative ethical questions thus, if we accept the normative interpretation, are inseparable from the conditions of responsibility.

I would like to emphasize that this thought does not involve any circularity. Moreover, this thesis is less radical than it might seem for the first sight. Remember, our discussion centers around a moral principle, the Control Principle, which, I assumed, has a direct impact on the conditions of moral responsibility. Thus, it comes as no surprise that substantive moral principles have a bearing on theories of moral responsibility. Obviously, talking about principles of fairness is different from being committed to a normative ethical theory as a whole. However, we can find several examples also for this latter case. For instance, when Wallace argues that it is fair to blame someone only if she has violated a moral obligation, he seems to commit himself to some sort of Kantianism, very broadly understood. Or, more controversially, when Watson characterizes responsibility as attributability as being open to so-called areatic appraisals about one's identity as expressed in her values and ends, he leans toward an Aristotelian framework.[30] And probably it is needless to explain how Smart's theory of responsibility is intertwined with his consequentialist commitments.

This methodology, however, often runs into serious problems, which we have already encountered when discussing Smith's criticism of Wallace's account in Chapter 7. To briefly recall it, Smith argues that there are moral considerations which regulate the fairness of holding someone responsible (in the sense of expressing reactive attitude and sanctioning behavior), but which, at least intuitively, have no bearing on the agent's responsibility itself. Smith names three such considerations: the moral standing of the blamer, the seriousness of the fault and the agent's response to her own wrongdoing. If the blamer is guilty of the very same moral fault as the one blamed; if the wrongdoing is relatively minor; or the agent already reproaches herself for it, then it often seems inappropriate to blame her. However, these considerations seem "external" to issues of responsibility.

There I argued that although these counterintuitive results indeed pose difficulties, these do not threaten the normative interpretation itself, but Wallace's characterization of holding responsible. In the present context, however, we cannot dismiss this criticism anymore. In the previous paragraph I claimed that normative ethical considerations have to be taken into account in any theory of responsibility. But now it seems that only *some* normative issues

[30] Watson himself makes an explicit move to this direction, when he argues that in the third book of *Nicomachean Ethics* Aristotle defines the attributability (as opposed to accountability) conditions of responsibility. See p. 273.

should be considered, while others, according to our intuitive judgments, should be ignored. But how could we discriminate between them in a principled way?

To illustrate the gravity of the problem, let's take on Smith's first point: the question of *entitlement* to blame. There are many cases where the agent's responsibility is undisputed but there is no one who could legitimately address sanction or reward, because addressing sanction and reward (or even *expressing* moral criticism) arguably requires some license or moral stand which some might lack. As Adams puts it: "Whether you have a right or duty to reproach another person may depend on your stake in the matter, or your relation to him, or on whether he has (explicitly or implicitly) invited you to correct him." (Adams 1985, p. 23) But no one would say that the absence of someone entitled to hold the agent responsible would make the agent any less responsible for what she did, omitted, felt or thought.

Obviously, one can say plenty of things in defense of Wallace. The most obvious (although somewhat question-begging) answer is to say that our initial definition should indeed be supplemented in the following way: one is morally responsible for something if and only if it would be fair for someone with appropriate moral standing to blame her. This is not an easy way out, however: beside the ungrateful job of defining what we mean by "appropriate moral standing" we should also give an informative and hopefully non-circular, principled explanation why moral standing is needed in order pick out responsible agency. Otherwise the revised definition seems *ad hoc*.

I see another, more promising way to answer Smith's objection, but it is beyond the scope of this chapter to argue for it in details. In short, my proposal is based on the recognition that when we talk about the fairness of holding responsible, we always occupy the recipient's perspective. The question is not whether an act of blame, for instance, is fair *tout court*, but whether it is fair *to the agent* to blame her[31]. Blame, ordinarily understood, might impose burdens on the agent, because it involves an impairment of relationship on the part of the blamer; it can morally discredit the blamed person in the eyes of third parties; question her moral integrity or force her to bear the burden of responsibility by feeling guilt, making apology or compensate for the harm done. The central question, as I have earlier emphasized, is whether it is fair (in general or in the particular case) to impose such burdens on the agent—and *this* question is arguably independent from the moral standing of the blamer. That is, it might be that it is unfair of me to blame someone for not having arrived in time, given that I am habitually late, but it is still fair to blame her in the responsibility-relevant sense (since, for instance, we think, in a retributivist fashion, that she deserves to feel bad about being late). That these two claims can be made compatible shows how heterogeneously we use the term "fairness" and consequently how much caution it requires to discriminate between the different concepts.

I cannot go on and elaborate on these claims. Neither can I defend Wallace's account from all of Smith's objections. I doubt that there would be a uniform way to rebut all these objections. Rather, a piecemeal approach seems desirable: to examine and explain all the exceptions one by one.

However, a general moral can be drawn from these cases: as we start considering more and more "serious" forms of responsibility-attribution, we will run into more and more such cases where judgments about the proper moral response and judgments of responsibility intuitively come apart. This becomes particularly clear if we briefly explore an extremely serious form of responsibility-attribution: the case of imprisonment. There is a variety of considerations influencing the appropriateness of holding someone responsible in this sense. These

[31] We have already touched upon this point when discussing Hieronymi's way of handling the charge of unfairness in Chapter 8.

considerations can be forward or backward-looking, and can raise issues of retributive, corrective or distributive justice[32]. For instance, few would deny that the moral justifiability of imprisonment depends partly on its preventive potential or its efficacy in rehabilitation. These considerations are, however, obviously independent from judgments about responsibility.[33] Thus, if one were to define responsibility as "being the fair target of imprisonment on the basis of it", then predictably she would have to face dozens of objections structurally similar to the ones Smith raised: it will seem that the question of proper moral response is largely independent from the question whether the agent is morally responsible for her action.

In this respect attributionist accounts have an obvious advantage. If imprisonment is on one side of a spectrum, then the attributionist concept of moral appraisal, understood as an unexpressed judgment about the agent's reasons or evaluative judgments, is on the opposite side. Theories of responsibility, according to attributionism, only have to define those minimal conditions which make some action, omission or attitude a proper object of moral appraisal by virtue of its being attributable to the agent. Since their characterization of responsibility-attribution hardly raises *any* moral consideration about fairness, they can easily avoid those problems which Smith raised against Wallace's account. According to attributionists, to justify any stronger moral response (such as reactive attitudes, sanctions and rewards) to a particular instance of responsible agency we need either another concept of responsibility (such as substantive responsibility in the case of Scanlon or accountability in the case of Watson) or we have to take into account such moral and ethical considerations which are completely independent of one's responsibility.

[32] Here I do not want to take a stand on the question whether these are really independent concepts of justice or not.

[33] One could object that this is because I conflate *legal* and *moral r*esponsibility. I disagree: the justifiability of imprisonment is a *par excellence* moral issue, which does not necessarily presuppose any notion of legal responsibility.

One of the advantages of this approach, as I emphasized it before, is that it widens the scope of responsibility so that it conforms better to our ordinary judgments of responsibility than other, more restricted accounts. But there is also another one which comes from the relative independence of issues of responsibility (i) and issues of the proper moral response (iii). I find that if we want to justify sanctions and rewards by virtue of the agent being responsible for the given action and attitude, then we make our theory of punishment dependent on our theory of responsibility. If, for instance, we build a desert-based account of moral responsibility and this theory serves to justify sanctioning the agent for her wrongdoings, then it will be difficult to explain why we allow for, for instance, consequentialist considerations in determining the moral desirability of punishment. If, by contrast, we develop a consequentialist account of moral responsibility, then we have to face all the familiar and severe challenges which consequentialist normative theories normally do with regard to punishment. The acceptance of the attributionist framework makes it possible to handle these problems independently and thus develop a theory of punishment as sophisticated, mixed and nuanced as the problem of punishment really is.

At this point of the discussion it is probably needless to say that Thomas Scanlon's work in *What We Owe to Each Other* is one of the most prominent examples of the constant interplay between "purely" normative questions and theories of responsibility. We already encountered in the Chapter 1 his concept of substantive responsibility, which is meant to fix questions about distributive and compensatory duties as they arise in social cooperation. Also, we could then see how substantive moral principles—such as Scanlon's account of the value of choice or the so-called Forfeiture View which he rejects—enter into his discussions about moral responsibility. Later on, in the previous chapter I presented his rejection of the Desert Thesis, illustrating how seemingly independent moral considerations affect our theory of responsibility. To finish the discussion, now I would like to raise yet another methodological issue, which might illuminate how inextricably these two spheres are intertwined.

As I previously suggested, the task of determining the necessary and sufficient conditions of responsibility is traditionally conceived as being in some sense prior to the job of our normative theories determining moral standards and duties[34]. This conception does not entail complete independence between the two spheres. Rather, the traditional division of labor would look like this: those providing a theory of the conditions of responsibility have to determine which sphere of human agency moral standards can be justifiably applied to. Moral standards, in turn, have to comply with theories of responsibility as far as they can be applied only to the sphere which a given theory of responsibility determines. It is as if theories of responsibility defined the territory of normative theories, but within those boundaries every decision is in the hand of the normative theorist.

Although I discussed in length attributionists' non-standard approach to several traditional issues of the free will debate, in this respect I seemed to presuppose that this picture is correct. So far we talked about judgments of reasons as such privileged parts of human psychology which entitle us to make further assessments about the agent. Also, I assumed that the reason why judgment-sensitive attitudes are the direct objects of responsibility-attribution is that they reflect the rational activity of the agent and thus, in some sense, they are up to us. Judgments and consequently judgment-sensitive attitudes *belong to us*—this is why we can be assessed on the basis of them.

I would like to argue that Scanlon uses a different methodology. This alternative reading of his attributionist theory would say that the reason why we can hold people responsible for

[34] Here I deliberately ignore issues of rightness and wrongness, since, according to the common view, judgments about right and wrong can be maintained even if we are not morally responsible for anything.

their judgment-sensitive attitudes (or for the lack of them) is that by having or lacking them they fell below some standard of rational and moral self-governance. It might seem that this way of putting things does not add much to what has already been told. But, quite on the contrary, it forces us to reconsider the status of judgments in Scanlon's attributionist framework. It seems that judgments are crucial only as far as they are constitutive of rational and moral self-governance which in turn is what moral standards can be applied to. And the reason why judgment-insensitivity exempts agents from responsibility is simply the fact that judgment-insensitive attitudes do not take part in rational and moral self-governance, and thus they are not subject to moral and rational standards. It is the direction of explanation which has been reversed: it is not that free and responsible agency defines the region which moral standards can be applied to, but that the standards of rational and moral behavior define the target of moral and rational assessment and thus the scope of responsibility.

I find that this later interpretation of attributionism fits much better with the project Scanlon works on in *What we owe to each other* than the previous, widely accepted one. First and foremost, this understanding of the problem of moral responsibility generates new and different requirements for a theory of responsibility. The main question is not anymore "Who is responsible for what kind of things on what basis?", but "What makes those standards of rational and moral behavior justified, which will define the target and content of moral assessment?" And this is the main topic of Scanlon's book.

Also, it is worth noting that I already took the first step toward this interpretation in the previous chapter, when I argued that attributionist accounts are interested rather in the reasonableness of moral demands than in defining the standard excusing and exempting conditions. Scanlon goes one step further, because he does not separate the reasonableness of the demand from other moral and rational considerations which bear on the desirability and justifiability of certain standards. Scanlon’s central question in *What We Ought To Each Other* is what standards and duties we should accept and this “should” includes several other considerations than reasonableness. His theory of responsibility is inseparably bound to his substantive moral theory.

Conclusion

If a theory is non-standard in several important aspects, then it is reasonable to suppose that these different aspects are somehow related to each other—this was the idea which led me while exploring attributionist theories of moral responsibility. Attributionism, as presented by Thomas Scanlon and Angela Smith, is silent about significant issues of the literature on free will and moral responsibility and denies claims which are often taken to be self-evident or obvious.

According to Neil Levy's formulation, "on the attributionist account, I am responsible for my attitudes, and my acts and omissions insofar as they express my attitudes, in all cases in which my attributes express my identity as a practical agent. Attitudes are thus expressive of who I am if they belong to the class of *judgment-sensitive attitudes*" (Levy 2005). Judgment-sensitive attitudes include, among other things, beliefs, emotions and intentions, but also spontaneous reactions such as noticing something or caring about somebody.

For attributionist accounts attitudes are the basic objects of responsibility-attribution, the things for which we are directly responsible. Oddly enough, from attributionism it follows that we are only indirectly responsible for our actions—we are responsible for them only as far as they are expressions of our judgment-sensitive attitudes. This is one of the reasons why attributionist accounts explicitly deny the importance which is traditionally attributed to voluntariness, choice and consciousness in establishing the agent's responsibility. In my thesis I gave a comprehensive analysis of attributionist accounts, highlighted the advantages of their solutions and discussed all the salient critical points which have been raised against them. Although I did not attempt to defend attributionist accounts from every criticism, I could hopefully convince some readers that attributionism has several appeals which make it a genuine rival of more traditional accounts of moral responsibility.

I find attributionism especially attractive because it touches upon such fascinating and confusing problems of responsibility-attribution which are relatively rarely discussed and even less often adequately explained. By taking on the issue of moral responsibility for emotions, beliefs and traits, on the one hand, and responsibility for carelessness, forgetfulness and negligence, on the other hand, theories of responsibility have to face special challenges.

I argued that the central difficulty with these so-called problematic cases is that they seem to contradict to an intuitively appealing but yet unanalyzed moral principle, the Control Principle, which states that it is unfair to hold people responsible for such things which are beyond our control. Much of what has been told in my thesis can be seen as different strategies to dissolve the following inconsistent triad:

> P1: It is unfair to hold people responsible for things over which they do not exercise control. (Control Principle)
>
> P2: In problematic cases of responsibility the agents do not exercise control.
>
> P3: In problematic cases the agents are morally responsible.

Thus, it seems that ordinary judgments of responsibility regularly go against a powerful moral principle, which none of us would like to reject, especially not without further argumentation. In Chapter 2 I gave an abstract mapping of the possible strategies to eliminate this tension, although later discussions revealed newer and newer methods.

Interestingly, even attributionists seem to diverge in this respect. Thomas Scanlon argues against P1 by providing an unusually weak concept of blame and moral criticism, which can be fair even if the agent did not exercise control—this is the strategy which I discussed in Chapter 8. Angela Smith, by contrast, denies an implicit premise which contributes to the inconsistent triad, i.e., that one is morally responsible for something, if and only if it is fair to hold her responsible on the basis of that thing. This claim, which is generally assumed by theorists of the field, was put forward by R. J. Wallace who called it the normative interpretation of moral responsibility. In Chapter 7 I defended Wallace's view in length, arguing that Smith's criticisms fail to undermine it. Also, I argued that the normative interpretation should be understood as a general schema which has far-reaching methodological consequences. Probably the most important conclusion is that in order to determine the conditions of moral responsibility first we have to fix three variables: the scope of responsibility-attribution, the nature of the relevant responsibility-attributing practices and the moral considerations about fairness which should be applied.

Appreciating the consequences of the normative interpretation also helps us to understand the core strategy which attributionist theories follow: by providing a relatively weak concept of moral appraisal (this is true also in Smith's case) they aim to eliminate concerns about the fairness of holding responsible. By using this method they open up the way to judgments of responsibility with a much wider scope than what is usually supposed.

Also, the discussion of the normative interpretation hopefully made it clear that certain moral considerations inevitably take part in determining the conditions of responsibility. In exactly what form and to which extent is this interplay inevitable? I have only tentative answers to this question, which I presented in the second half of my thesis, especially in Chapter 11. At the very extreme, some theories of responsibility might seem as simple subsets of normative ethical theories of rightness and wrongness, standards and obligations. Since this topic is seriously underexplored in the relevant literature, I can only hope that my first, cautious attempts of clarification and conceptualization will help to step forward in the understanding of these fascinating issues.

Bibliography

Adams, R. M. 1985. Involuntary sins. *The Philosophical Review* 94, pp. 3–31.

Aristotle. 2002. *Nicomachean Ethics*. Translated by Christopher Rowe. With philosophical introduction and commentary by Sarah Broadie. Oxford: Oxford University Press.

Austin, J. 1956. A Plea for Excuses. *Proceedings of the Aristotelian Society* 57, pp. 1--30.

Ben-Ze'ev, A. 1997. Emotions and Morality. *The Journal of Value Inquiry* 31, pp. 195–212.

Ben-Ze'ev, A. 2000. *The Subtlety of Emotions.* MIT Press.

Bok, H. 1998. *Freedom and Responsibility*. Princeton University Press.

Bomann-Larsen, L. 2010. Revisionism and Desert. *Criminal Law and Philosophy* 4, pp. 1–16.

Brandt, R. 1969. A Utilitarian Theory of Excuses. *The Philosophical Review* 78, pp. 337–361.

Clarke, R. (n.d.) Moral Responsibility as Appraisability. Unpublished manuscript.

Clarke, R. 2009. Dispositions, Abilities to Act, and Free Will: The New Dispositionalism. *Mind* 118, pp. 323–351.

D'Arms, J. & Jacobson, D. 2003. The significance of recalcitrant emotion (or, anti-quasijudgmentalism). In Anthony Hatzimoysis (ed.), *Philosophy and the Emotions*, pp. 127–145. New York: Cambdridge University Press.

D'Arms, J. & Jacobson, D. 2000. The Moralistic Fallacy: On the "Appropriateness" of Emotions. *Philosophy and Phenomenological Research* 61, pp. 65–90.

Davidson, D. 1963. Actions, Reasons, and Causes. *Journal of Philosophy* 60, pp. 685–700.

Dennett, D. C. 1984. Elbow room: the varieties of free will worth wanting. Cambridge, Mass.: MIT Press.

Dworkin, G. 2011. *Justice for Hedgehogs*. Cambridge Mass.: Belknap Press of Harvard University Press.

Fara, M. 2008. Masked Abilities and Compatibilism. *Mind* 117, pp. 843–865.

Finlay, S. & Schroeder, M. 2008. Reasons for Action: Internal vs. External. Retrieved May 1, 2013 from http://plato.stanford.edu/entries/reasons-internal-external/

Fischer, J. M. & Togazzini, N. 2009. The Truth about Tracing. *Noûs* 43, pp. 531–556.

Fischer, J. M. & Togazzini, N. 2011. The Physiognomy of Responsibility. *Philosophy and Phenomenological Research* 82, pp. 381–417.

Fischer, J. M. & Ravizza, M. 1998. *Responsibility and Control: a Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, J. M. 2012a. *Deep Control: Essays on Free Will and Value*, Chapter 8. New York: Oxford University Press.

Fischer, J. M. 2012b. Responsibility and autonomy: The problem of mission creep. *Philosophical Issues* 22, pp. 165–184.

Frankfurt, H. 1969. Alternate Possibilities and Moral Responsibility. Journal of Philosophy 66, pp. 829–39.

Frankfurt, H. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68, pp. 5–20.

Frankfurt, H. 1976. Identification and Externality. Reprinted in *The Importance of What We Care About*, pp. 58–69. New York: Cambridge University Press, 1998.

Frankfurt, H. 1987. Identification and Wholeheartedness. Reprinted in *The Importance of What We Care About*, pp. 159–177. New York: Cambridge University Press, 1998.

Goldie, P. 2004. *On Personality*. London: Routledge.

Haji, I. 1999. Moral Anchors and Control. *Canadian Journal of Philosophy* 29, pp. 175–203.

Haji, I. 2002. Compatibilist Views of Freedom and Responsibility. In Robert Kane (ed.), *The Oxford Handbook of Free Will*, pp. 202–228. Oxford: Oxford University Press.

Hieronymi, P. 2004. The Force and Fairness of Blame. *Philosophical Perspectives* 18, pp. 115–48.

Hieronymi, P. 2006. Controlling Attitudes. *Pacific Philosophical Quarterly* 87, pp. 45–74.

Hieronymi, P. 2008. Responsibility for Believing. *Synthese* 161, pp. 357–373.

Hieronymi, P. 2009. Two Kinds of Agency. In O'Brien, L. & Soteriou, M. (eds.), *Mental Actions*, pp. 138–162. Oxford University Press.

Huoranszki, F. 2011. Freedom of the Will: A Conditional Analysis. New York: Routledge.

Kane, R. 2002. Free Will: New Directions for an Ancient Problem. In: Kane (ed.), *Free Will*, pp. 222–248. Blackwell.

Kane, R. 1996. *The Significance of Free Will*. New York: Oxford University Press.

King, M. 2009. The Problem with Negligence. *Social Theory and Practice* 35, pp. 577–595.

King, M. 2012. Moral Responsibility and Merit. *Journal of Ethics & Social Philosophy* 6, pp. 1–17.

Levy, N. & McKenna, M. 2009. Recent Work on Free Will and Moral Responsibility. *Philosophy Compass* 4/1, pp. 96–133.

Levy, N. 2005. The Good, the Bad and the Blameworthy. *Journal of Ethics & Social Philosophy* 1, no. 2.

Levy, N. 2008. Restoring Control: Comments on George Sher. *Philosophia* 36, pp. 213–221.

Mason. M. 2011. Blame: Taking it Seriously. *Philosophy and Phenomenological Research* 83, pp. 473–481.

McCormick, M. 2011. Taking control of belief. *Philosophical Explorations* 14, pp. 169–183.

McKenna, M. 2008. Putting the lie on the control condition for moral responsibility. *Philosophical Studies* 139, pp. 29–37.

McKenna, M. 2009. Compatibilism. Retrieved May 1, 2013 from http://plato.stanford.edu/entries/compatibilism/

Nagel, T. 1979. Moral luck. In *Mortal Questions*, 24–38. Cambridge: Cambridge University Press.

Nelkin, D. 2013. Moral luck. Retrieved May 1, 2013 from http://plato.stanford.edu/entries/moral-luck/

Nelkin, D. 2009. Responsibility, rational abilities, and two kinds of fairness arguments. Philosophical Explorations 12, pp. 151–165.

O’Brien, L. & Soteriou, M. (eds.) 2009. *Mental Actions*. Oxford University Press.

Owens, D. 2000. *Reason without Freedom*. London: Routledge.

Owens, D. 2003. Does Belief Have an Aim? *Philosophical Studies* 115, pp. 283–305.

Pereboom, D. 1995. *Determinism* al Dente. *Noûs* 29, pp. 21–45.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Pereboom, D. 2008. Defending Hard Incompatibilism Again. In Trakakis, N. and D. Cohen (eds.), *Essays on Free Will and Moral Responsibility*, pp. 1–34. Cambridge Scholars Publishing: Newcastle upon Tyne.

Russel, P. 1992. Strawson's Way of Naturalizing Responsibility. *Ethics* 102, pp. 287–302.

Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Scanlon, T. 2008. *Moral dimensions: permissibility, meaning, blame*. Cambridge: Harvard University Press.

Schlick, M. 1939. When is a Man Responsible. In *Problems of Ethics*, 141–158. New York: Prentice-Hall.

Schlossberger, E. 1986. Why We Are Responsible for Our Emotions. *Mind* 377, pp. 37–56.

Sher, G. 2001. Blame for Traits. *Noûs* 35, pp. 146–161.

Sher, G. 2005. Kantian Fairness. *Philosophical Issues* 15, pp. 179–192.

Sher, G. 2006. Out of Control. *Ethics* 116, pp. 285–301.

Sher, G. 2009. *Who Knew?* New York: Oxford University Press.

Shoemaker, D. 2011. Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics* 121, pp. 602–632.

Smart, J. J. C. 1961. Free-Will, Praise and Blame. *Mind* 70, pp. 291–306.

Smilansky, S. 2002. Free Will, Fundamental Dualism, and the Centrality of Illusion. In Robert H. Kane (ed.), *The Oxford Handbook of Free Will*, pp. 425–442. New York: Oxford University Press.

Smilansky, S. 2008. Free Will and Fairness. In: Nick Trakakis & Daniel Cohen (eds.), *Essays on Free Will and Moral Responsibility*, pp. 235–247. Newcastle upon Tyne: Cambridge Scholars Publishing.

Smith, A. M. 2005. Responsibility for Attitudes. *Ethics* 115, pp. 236–271.

Smith, A. M. 2008a. Control, responsibility, and moral assessment. *Philosophical Studies* 138, pp. 367–392.

Smith, A. M. 2008b. On Being Responsible and Holding Responsible. *The Journal of Ethics* 11, pp. 465–484.

Smith, A. M. 2012. Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics* 122, pp. 575–589.

Smith, H. M. 1983. Culpable Ignorance. *Philosophical Review* 92, pp. 543–571.

Smith, H. M. 2011. Non-Tracing Cases of Culpable Ignorance. *Criminal Law and Philosophy* 5, pp. 115–146.

Smith, M. 2003. Rational Capacities. In Stroud and Tappolet (eds.), *Weakness of Will and Practical Irrationality*, pp. 17–38. Oxford: Oxford University Press.

Sneddon, A. 2005. Moral Responsibility: The Difference of Strawson, and the Difference It Should Make. *Ethical Theory and Moral Practice* 8, pp. 239–264.

Solomon, R. C. 1973. Emotions and Choice. *Review of Metaphysics* 27, pp. 20–41.

Strawson, G. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

Strawson, G. 1994. The Impossibility of Moral Responsibility. *Philosophical Studies* 75, pp. 5–24.

Strawson, P. F. 1962. Freedom and Resentment. *Proceedings of the British Academy* 1962, pp. 1–25.

Szigeti, A. 2012. Revisiting Strawsonian Arguments from Inescapability. *Philosophica* 85, pp. 91−121.

Thomson, J. J. 1993. Morality and Bad Luck. In D. Statman (ed.), *Moral Luck*, pp. 195–217. Albany: State University of New York Press.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.

Vargas, M. 2004. Responsibility and the Aims of Theory: Strawson and Revisionism. *Pacific Philosophical Quarterly* 85, pp. 218–241.

Vargas, M. 2005. The Trouble with Tracing. *Midwest Studies in Philosophy* 29, pp. 269–291.

Velleman, D. 2000. On the Aim of Belief. In David Velleman (ed.), *The Possibility of Practical Reason*, pp. 244–283. Oxford University Press.

Vihvelin, K. 2004. Free Will Demystified: A Dispositionalist Account. *Philosophical Topics* 32, pp. 427–450.

Walker, M. U. 1991. Moral Luck and the Virtues of Impure Agency. *Metaphilosophy* 22, pp. 14–27.

Wallace, R. J. 1995. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Wallace, R. J. 2011. Internalism about Responsibility. Unpublished manuscript.

Watson, G. 1975. Free Agency. *Journal of Philosophy* 72, pp. 205–220.

Watson, G. 1987. Free Action and Free Will. *Mind* 96, pp. 145–172.

Watson, G. 1987. Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In: Schoeman, F. (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, pp. 256–286. Cambridge: Cambridge University Press.

Watson, G. 1998. Some Worries about Semi-Compatibilism: Remarks on John Fischer's *The Metaphysics of Free Will. Journal of Social Philosophy* 29, pp. 135–143.

Watson, G. 1996. Two Faces of Responsibility. *Philosophical Topics* 24, pp. 227–248. Reprinted in *Agency and Answerability*, pp. 260–288. Oxford: Clarendon Press, 2004.

Whiting, D. 2012. Does Belief Aim (Only) at the Truth? *Pacific Philosophical Quarterly* 93, pp. 279–300.

Williams, B. 1979. Internal and External Reasons. Reprinted in *Moral Luck*, pp. 101–13. Cambridge: Cambridge University Press, 1981.

Williams, B. 1981. Moral luck. In: *Moral Luck*, pp. 20–39. New York: Cambridge University Press.

Wolf, S. 1990. *Freedom within Reason*. Oxford: Oxford University Press.