UNIVERSITY OF READING

MORAL RESPONSIBILITY AND SUBVERTING CAUSES

A thesis submitted for the degree of Doctor of Philosophy

by Andy Taylor

Department of Philosophy

September 2010

Abstract

I argue against two of the most influential contemporary theories of moral responsibility: those of Harry Frankfurt and John Martin Fischer. Both propose conditions which are supposed to be sufficient for direct moral responsibility for actions. (By the term "direct" moral responsibility, I mean moral responsibility which is not *traced* from an earlier action.) Frankfurt proposes a condition of "identification"; Fischer, writing with Mark Ravizza, proposes conditions for "guidance control". I argue, using counterexamples, that neither is sufficient for direct moral responsibility.

My counterexample cases are based on recent research in psychology which reveals many surprising causes of our actions. Some of this research comes from the field of "situationist" social psychology; some from experiments which reveal the influence of automatic processes in our actions. Broadly, I call such causes "subverting" when the agent would not identify with her action, if she knew all the causes of the action. When an action has subverting causes, the agent is not directly morally responsible for it, even though she may meet the conditions specified by Frankfurt and Fischer.

I also criticise the theories of Eddy Nahmias and John Doris, who have both engaged specifically with the threats posed to moral responsibility by situationist research. Against Doris and Nahmias, I argue that their conditions are neither necessary nor sufficient for direct moral responsibility.

My final objective is to argue that there are many everyday actions for which we mistakenly hold agents morally responsible. I review evidence that there are many everyday actions which have subverting causes. Many of those are actions for which we currently hold agents morally responsible. But I argue that, in many of those same actions, the agents are not in fact morally responsible – they bear neither direct nor traced moral responsibility.

Table of Contents

Acknowledgements

For financial sponsorship I thank and acknowledge the support of the Arts and Humanities Research Council (AHRC).

I'm grateful to the staff and students of the University of Reading for my training in philosophy, and the opportunity to develop and discuss these ideas.

I thank my family and friends for their patience and understanding while I've been working on this thesis.

I am deeply grateful to my supervisors, Galen Strawson and Bart Streumer, for invaluable ideas, advice and generous giving of their time at every stage of my work.

Finally, for her support in all of these areas and many more, I thank my wonderful wife Caroline.

1. INTRODUCTION

Samantha is on her way to give a presentation about the parable of 'The Good Samaritan'. She walks past a man who is slumped in a doorway apparently in distress, without offering to help. A cause of her action is that, a few moments ago, Samantha was told that she was running late for the presentation. If she had not been told that she was running late, Samantha would have stopped to help the person slumped in the doorway. If Samantha learned that being told she was running late was a cause of her walking past the slumped man, she would be shocked, and she would strongly regret her action. She would not *identify with*, nor *take responsibility for*, her action.[1]

Ingrid enters a room and interrupts a conversation which was already in progress between two other people. A cause of her action is that, a few minutes earlier, Ingrid was completing a word puzzle. Some of the words in the puzzle were connected with a theme of rudeness. If she had been completing a word puzzle with a neutral theme, Ingrid would not have interrupted the conversation. If Ingrid learned that completing the word puzzle was a cause of her interrupting the conversation, she would be shocked, and she would regret her action. She would not *identify with*, nor *take responsibility for*, her action.

These are two examples of phenomena which I call "subverting causes" of actions. Broadly, causes are "subverting" when the agent would not identify with her action, if she knew all the causes of the action.[2] Recent and ongoing research in psychology is steadily uncovering more and more examples of similar causes of actions. It's my opinion that subverting causes are common as causes of everyday actions, and that philosophers have been slow to recognise the threat which they pose to moral responsibility.

Using examples of actions with subverting causes, I will argue against two of the most influential contemporary theories of moral responsibility – those of Harry Frankfurt and John Martin Fischer. Both propose conditions which are supposed to be sufficient for direct moral responsibility for actions. (By the term "direct" moral responsibility, I mean moral responsibility which is not *traced* from an earlier action. I'll discuss tracing of moral responsibility in section 1.2 below.) Frankfurt proposes a condition of "identification"; Fischer, writing with Mark Ravizza, proposes conditions for "guidance control". I will argue that actions with subverting causes provide counterexamples to both theories: an agent may meet the conditions proposed by each theory and yet not be directly morally responsible for her action. I'll call this my "main argument"; it has the same basic structure when made against each of the two theories.

My argument against Frankfurt's condition employs important elements of his own position. I'll argue that, after recognising that some actions do have subverting causes, followers of Frankfurt should accept that his identification condition is not sufficient for direct moral responsibility. Similarly, my argument against Fischer and Ravizza draws on important elements of their theory. I'll argue that, recognising the phenomena of subverting causes, followers of Fischer and Ravizza should agree that their conditions for guidance control are not sufficient for direct moral responsibility.

I also have a number of secondary objectives. I'll make a positive proposal of a new condition which I think is necessary for direct moral responsibility. It incorporates the central elements of Frankfurt's theory, and I'll argue that his followers should accept my proposed condition.

I'll also discuss two philosophers who have engaged specifically with the threats to responsibility posed by research in "situationist" social psychology, which has revealed

[1] I'll explain the italicised phrases in chapters 2 and 6 respectively.
[2] I'll discuss subverting causes in much more detail in chapter 3.

subverting causes of actions like Samantha's above.[3] Eddy Nahmias and John Doris have both proposed conditions of direct moral responsibility which draw on identificationist theories like Frankfurt's, but include revisions to take account of actions with situationist causes. I will criticise and reject both Doris's and Nahmias's accounts. I'll argue against each theory that its conditions are neither necessary nor sufficient for direct moral responsibility.

My final objective is to argue that there are many everyday actions for which we mistakenly hold agents morally responsible. In what I'll call my "sceptical argument", I'll draw on evidence that there are many everyday actions which have subverting causes. Many of those are actions for which we currently hold agents morally responsible. But, in many of those same actions, the agents are not in fact morally responsible – they bear neither direct nor traced moral responsibility.

Though they share a common theme, each of my arguments against a particular philosopher's position is independent of the others. For example if my argument against Fischer and Ravizza is shown to be unsuccessful, my argument against Frankfurt will be unaffected. My final "sceptical argument" begins with the premise that agents are not directly morally responsible for actions which have subverting causes. It therefore requires that one of the previous arguments has been successful. However, if my sceptical argument fails, the previous arguments are unaffected.

1.1. SCOPE AND ASSUMPTIONS

1.1.1. Scope

The scope of my enquiry is limited to *an agent's moral responsibility for actions which he has performed*. In this section I'll clarify that statement by pointing out some related areas of enquiry which fall outside my scope. Then, in the following two sections I'll examine the central concepts of moral responsibility and actions.

I will exclude several interesting areas of enquiry from my scope. For the most part I do so merely to help simplify my arguments, and not because I think those arguments aren't applicable in other areas. My aim is to make my arguments as clear and precise as possible within a relatively narrow scope of enquiry. If those arguments are successful, I anticipate that they will also have a wider scope of application with only minor modifications.[4]

I am concerned with actions performed by *adult human agents*. I won't be concerned with animals or inanimate agents, in any of the senses in which those may be agents. Nor will I be discussing actions performed by children. Unless stated otherwise, all the examples I'll discuss involve *normal* adults, by which I mean people who do not lack any normal physical or mental abilities.

I will focus on *moral responsibility* for actions, and not on the related concepts of free action, autonomous action, control or free will. There are several ways to interpret the relations between these concepts and moral responsibility, so for simplicity I'm going to set aside those terms.[5]

I will concentrate on moral responsibility for *actions which are performed*. I won't be addressing responsibility for omissions – actions which are not performed. Nor will I discuss

[3] My definition of "subverting causes", given in section 3.1, encompasses but is not restricted to "situationist" causes. In chapter 3 I'll discuss a range of evidence from research in psychology, including but not restricted to situationist research, which suggests that subverting causes are common.

[4] For example, I think my arguments could be adapted to cover moral responsibility for omissions as well as actions, but I will not argue for that here.

[5] I will briefly discuss the relation between free will and moral responsibility in Frankfurt's theory, and the relation between control and moral responsibility in Fischer and Ravizza's theory. But I will focus on moral responsibility in my own arguments.

responsibility for any other kinds of events, or for states of affairs. For example, suppose Stan puts his dirty coffee mug down on the kitchen table. Stan's wife may hold him responsible for *the table being stained* (a consequence of his action), or for *failing to clean up the stain* (an omitted action). I won't discuss these kinds of responsibility.

I will only address an agent's moral responsibility *for actions which he performed*. This excludes *vicarious* moral responsibility inherited by one person from another agent's action: for example, when a company director is vicariously morally responsible for an employee's action.

1.1.2. Actions

In this section I'll briefly set out the working assumptions I'll make about actions. There are many difficult issues in the philosophy of action with which I cannot hope to engage in any detail. Again my objective is to argue as precisely as possible for conclusions which apply to a limited but very important set of actions. If I succeed then those conclusions will be significant; it may also be possible to extend my arguments, with minor modifications, to a wider set of actions.

I will usually discuss examples of *physical movements*, and not "mental actions" such as decisions or calculations performed by thought alone. But I do not claim that all actions must be physical movements.

I am concerned only with *intentional actions*. There are several competing philosophical theories about how to define an intentional action, but for my purposes I need not commit myself to any particular view. Fischer and Ravizza have very little to say about the nature of intentional action, but they appear to advocate a broadly Davidsonian causal theory.[6] Frankfurt on the other hand criticises causal theories of action in his 1978 paper 'The Problem of Action'. Nevertheless this disagreement has no important implications for my arguments. The examples I'll discuss will involve actions which are regarded as intentional on all the mainstream views.

Another contentious problem on which I will remain neutral concerns the individuation of actions. Neither Frankfurt nor Fischer has anything to say (as far as I'm aware) about how actions are to be individuated. I think the problem is orthogonal to the arguments I will be making: my arguments will be equally compelling on any of the mainstream theories of action individuation.[7]

1.1.3. Moral responsibility for actions

There is more than one view among philosophers about what it means to ascribe moral responsibility to agents.[8] In my two main arguments I criticise the theories of Frankfurt and Fischer and Ravizza. Ideally, when arguing against Frankfurt I would like to assume that his view of the concept is correct, and when arguing against Fischer and Ravizza that their concept of moral responsibility is correct. Unfortunately, I'm not aware that Frankfurt has explained his view of the concept of moral responsibility. Therefore I shall simply describe Fischer and Ravizza's account of the concept, and assume throughout that it is correct. I don't think my arguments against Frankfurt are greatly undermined by the lack of a statement from him on this point, for several reasons. It's likely that Frankfurt's view is quite similar to Fischer and Ravizza's, since they occupy broadly similar compatibilist positions in the wider

[6] They say that non-intentional actions, such as seizures and tics, "are not caused (in an appropriate way) by beliefs and desires" (1998:82). See also Davidson 1963.

[7] I will ignore tricky cases on which different theories might give different verdicts about the number of actions involved.

[8] For a summary of these views, see for example Eshleman 2009.

philosophical debates about free will and moral responsibility.[9] Furthermore, my arguments are about the *conditions* of moral responsibility; those arguments are not very much affected even if in fact Frankfurt has a different view about the meaning of the concept. Moreover, the same lack of a statement from Frankfurt on this point has not hampered many other criticisms of his account of the conditions for moral responsibility.

Fischer and Ravizza's definition of the meaning of the concept is as follows:

> being morally responsible is being *an appropriate candidate* for the reactive attitudes. A morally responsible agent is … a potential candidate for at least some of the reactive attitudes, but he need not be an *actual* recipient or target of any such attitude.[10]

The "reactive attitudes" are our natural responses to actions, including gratitude, indignation, respect, resentment, forgiveness and so on. These are reflected in practices such as punishing, blaming, condemning, and approving. These attitudes and practices are very closely connected with morally responsible agency, as Peter Strawson famously emphasised.[11] Fischer and Ravizza describe their view of the concept as a "Strawsonian view". Nevertheless they claim that their view diverges in an important way from Strawson's:

> Strawson's theory may reasonably be said to give an account of what it is for agents to be held responsible, but there seems to be a difference between being *held* responsible and actually *being* responsible … Strawson's theory risks blurring the difference between these two issues.[12]

Fischer and Ravizza argue (correctly, in my view) against Strawson that a practice of responsibility attribution which is the norm in a certain society can be evaluated from a standpoint external to that society.[13] They also quote approvingly an example from Jay Wallace in which a very charming man cheats and lies to a colleague. The wronged colleague "may have trouble working up any resentment or indignation", and yet it is clear that the liar is morally responsible: he is an appropriate candidate for the reactive attitudes in this case, despite not being an actual target of them.[14]

On Fischer and Ravizza's view, being *morally responsible* does not entail being either *blameworthy or praiseworthy*. They give the example of a bank teller who hands over money to a robber at gunpoint.[15] The teller may well act calmly and be morally responsible for this action.[16] Yet complying with the threat is very reasonable: we do not blame the teller for doing so. The teller can be morally responsible for handing over the money without being blameworthy for doing so. Furthermore, in Fischer and Ravizza's view, agents can be morally responsible for actions which are "morally neutral", and for which neither blame nor praise is appropriate.[17] However, in order to be blameworthy or praiseworthy for an action, an agent must be morally responsible for it.

---

[9] More specifically, Frankfurt shares Fischer and Ravizza's view that an agent can be morally responsible without being blameworthy for his action (1975:55-56).

[10] Fischer and Ravizza 1998:7. (Unless stated otherwise, italics in quoted passages are always taken from the original authors.)

[11] Strawson 1962.

[12] Fischer and Ravizza 1993:18. See also Fischer 1994:211-213.

[13] Fischer and Ravizza 1993:18.

[14] Wallace 1994:76-77; discussed in Fischer and Ravizza 1998:8.

[15] Fischer and Ravizza 1998:36n7.

[16] We might even say to the teller: "You acted very responsibly, in the circumstances" (G. Strawson 1986b:312).

[17] Fischer and Ravizza 1998:8n11 and 43.

A possible weakness in Fischer and Ravizza's position is that they do not explain *what it is to be an appropriate candidate* for the reactive attitudes.[18] However, I will simply assume that this view of the concept of moral responsibility is correct: my purpose is to argue that the *conditions* which they propose do not satisfy the requirements of the concept.

1.1.4. Determinism and compatibilism

I'll now briefly explain my stance on an issue which is crucial in many debates about moral responsibility. There are several ways to express the thesis called "determinism", but one of the simplest is this:

> the total set of facts about the past, together with the natural laws, entail all the facts about what happens in the present and future.[19]

Many people hold the view that determinism is incompatible with moral responsibility, in the sense that if determinism is true then moral responsibility is always impossible. That view is denied by all of the philosophers whose theories I'll be addressing, who are *compatibilists* about moral responsibility and determinism. They aim to specify the conditions in which agents can be morally responsible *even if* determinism is true.

My arguments do not presuppose any particular view about whether determinism is true, nor about whether it is compatible with moral responsibility. I will argue against certain compatibilist philosophers that they should accept my proposed revisions to their theories, *without* abandoning their compatibilist stance. I will mention determinism only fleetingly.

1.2. TRACED MORAL RESPONSIBILITY

Many philosophers hold that moral responsibility for an action need not be anchored in facts about the action and the agent at the time of acting. According to what I'll call "the tracing principle", an agent's moral responsibility can sometimes be "traced back" to facts about an *earlier* action and about the agent at that earlier time.[20] I'll use the term "traced moral responsibility" to refer to moral responsibility which is traced back to facts about an earlier time. In contrast, what I'll call "direct moral responsibility" is not traced back in this way: it is anchored in facts about the action and the agent at the time of acting.

An example will help to clarify the difference between these two types of responsibility. Suppose Trisha took a fix of a drug for the first time yesterday, knowing it to be extremely addictive. She now has a very strong craving for the drug, and takes another fix this morning. It may be that, on many theories of moral responsibility, Trisha does not meet the conditions for direct moral responsibility for her action of taking the drug this morning. She may be excused from direct moral responsibility on the grounds that she lacks the necessary control of her action; or that she could not have done otherwise, in some important sense; or that her desire to take the drug is not truly her own, because she wants not to have the craving to take the drug.

And yet many people would hold Trisha morally responsible for her action this morning, on the grounds that she took the drug yesterday, and that her doing so again this morning was reasonably foreseeable by her yesterday. It's natural to say that, in certain cases,

---

[18] It seems to me that an appeal to a related concept such as merit or desert is implicit but not acknowledged in this notion. Such an appeal seems to take their theory further away from Strawson's. Eshleman (2009) interprets Fischer and Ravizza's theory as a "merit-based" form of compatibilism.

[19] Fischer 2006:5.

[20] The tracing principle is usually extended to cover moral responsibility for *omissions* and *consequences* as well as actions, but I will consider it only in so far as it concerns responsibility for actions (in line with the scope of my overall enquiry).

moral responsibility for an action can be anchored in facts about an *earlier* action (or omission) and about the agent at that earlier time.[21]

My arguments in later chapters are about *direct* moral responsibility for actions. But it will often be important to retain a clear distinction between direct and traced moral responsibility. In particular, some of my replies to objections will draw on this distinction. In this section I'll introduce a generic and loosely defined version of the tracing principle, which should be acceptable to many of the philosophers whose theories I will discuss later. I'll then respond to an interesting criticism of the tracing principle made by Manuel Vargas, which raises a series of issues that will recur in later chapters.

I propose the following generic statement of the tracing principle, as it relates to moral responsibility for actions:

> An agent bears traced moral responsibility for action A1 if: she bears direct moral responsibility for an earlier action (or omission) A0; the later action A1 results from A0; and the resultant occurrence of A1 was reasonably foreseeable by the agent, at the time of A0.[22]

The agent's moral responsibility for action A1 can be "traced back" to facts about her earlier action (or omission) A0, or about her at that time.

In my terminology, it may well be the case that Trisha is not *directly* morally responsible for injecting the fix today. Her moral responsibility is not anchored in current facts about the action and about Trisha. But she does bear *traced* moral responsibility, because she did bear direct moral responsibility for her action of taking the drug yesterday, and the resultant occurrence of today's action was reasonably foreseeable by her when she did so.

Vargas discusses a set of cases which, he thinks, highlight a problem with the *epistemic* requirement in the tracing principle, which is captured in the phrase "reasonably foreseeable by the agent".[23] The most interesting of these cases involves Jeff, a manager who unreflectively makes an extremely insensitive comment to one of his employees just after he has made her redundant.[24] He is unreflective about his despicable treatment of this employee. Jeff is "a jerk": he often treats people rudely and insensitively, and is unreflective about doing so. When other people react poorly to his behaviour, he always writes it off as a shortcoming on their part. Jeff made deliberate and ultimately successful efforts to become a jerk at age 15, after noticing that other boys who were jerks received the most attention from girls.

Vargas makes three claims about the case. First, our intuitions tell us very clearly that Jeff is morally responsible for his insensitive comment. Second, Jeff does not bear *direct* moral responsibility for this action. Third, there is no suitable earlier action (or omission) to which his moral responsibility can be traced back, when the later action was reasonably foreseeable by Jeff.

---

[21] It seems very plausible that moral responsibility for an action can be traced from an earlier *omission*, as well as from an earlier action. For example, Trisha might bear traced moral responsibility for injecting her current fix if she had earlier omitted to take a less harmful substitute drug prescribed to reduce her craving.

[22] In this statement I draw on Vargas's phrasing (2005:269-274). In section 6.5 I'll discuss an alternative set of conditions for traced moral responsibility given by Fischer and Ravizza.

[23] This problem will be relevant to my arguments in chapters 5 and 6.

[24] Vargas 2005:271 and 275-276. I have changed the case very slightly; in my version there is a clearly identifiable single action for which Jeff may be held responsible. I have chosen to focus on this case among the four which Vargas raises because it poses the sternest challenge to the epistemic condition in the tracing principle. Fischer and Tognazzini (2009) also concentrate on this case, in their criticism of Vargas which I discuss below.

I agree with the first claim: we would hold Jeff morally responsible for making the insensitive comment. The second claim, on the other hand, looks somewhat implausible. Vargas seems to conclude that, because Jeff acts unreflectively from a long-standing character trait or habit, he is therefore not directly morally responsible for making his comment.[25] I don't think that follows.[26] However, I'll set aside this problem to examine the third claim. Assuming for the sake of argument that Jeff is *not* directly morally responsible for making the insensitive comment, can we trace his moral responsibility back to some earlier action? Vargas thinks not. At the time when Jeff freely decided, at age 15, to become a jerk – and he was directly morally responsible for doing so – he could not foresee that he would make an insensitive comment to an employee who had just been made redundant. Vargas concludes that either the tracing principle is flawed, or we must reject our strong intuition that an agent like Jeff is morally responsible for his comment.[27]

I think there are two objections which refute Vargas's argument. The first is made by Fischer and Tognazzini. They point out that Jeff's action could indeed have been foreseen by his younger self, when aged 15, *if it were specified more broadly*.[28] The action certainly could not have been foreseen when specified as "insulting that particular employee on that day after making her redundant as a result of becoming a jerk". Nor could it reasonably have been foreseen when specified as "one day insulting an employee after making her redundant as a result of becoming a jerk". But it could certainly have been foreseen when specified as "one day insulting someone as a result of becoming a jerk". The question now becomes: which of these must be reasonably foreseeable for Jeff to bear traced moral responsibility? The answer, according to Fischer and Tognazzini, is that the broadest is the most appropriate. Jeff bears traced moral responsibility for making his insensitive comment if "one day insulting someone" was reasonably foreseeable by him when he decided to become a jerk. I think this is a plausible claim, which reflects our practice of responsibility attribution.

The second objection I make to Vargas's argument is that there are more recent actions to which Jeff's moral responsibility for his comment can be traced. In making this point I continue (implausibly) to assume, for the sake of argument, that Jeff is not directly morally responsible for making his comment. Nevertheless it seems to me that even people who are jerks do occasionally reflect upon the impact of their habitual behaviour. Suppose for example that, a year ago, Jeff's last girlfriend accused him of habitual insensitivity before she left him, prompting Jeff to reflect over a brief period on whether his behaviour had indeed contributed to the failure of the relationship. At the end of that period he decided to do nothing to change his future behaviour.[29] I think it's plausible to say that his responsibility for the insensitive comment to his employee can be traced back to the moment a year earlier when he decided to do nothing to change his future behaviour.

Vargas seems to think that Jeff's decision to become a jerk at the age of 15 is the only prior moment in his history to which his responsibility could be traced.[30] I can see two possible reasons why Vargas might think this; I'm not sure which reason he has in mind, but I think neither is persuasive. The first is perhaps that responsibility can only be traced to the point at which Jeff positively or explicitly decides to become a jerk. The second may be that

[25] Vargas 2005:274-275.
[26] On this point I agree with Fischer and Tognazzini, who interpret Vargas's claim in the same way, and find it implausible (2009:535-536). They too set aside this criticism after a brief discussion, in order to concentrate on Vargas's third claim.
[27] Vargas 2005:271.
[28] Fischer and Tognazzini (2009:537).
[29] It doesn't matter whether we interpret this decision to do nothing as an action or as an omission, since responsibility for a later action can be traced back to an earlier action *or* omission. Equally, if Jeff simply decided nothing and did nothing after his brief period of reflection, this too is an omission to which his later responsibility can be traced.
[30] Vargas 2005:277.

Jeff is not directly morally responsible for any decisions or omissions affecting his insensitive behaviour at any point after he decided to become a jerk. The first reason would seem to contradict the tracing principle as I stated it above (in which I follow Vargas's own very loose definition). Moral responsibility can be traced to *any* action or omission from which the later action results. To defend the second possible reason, Vargas would presumably claim that there has been no point at which Jeff was *sufficiently reflective* about his behaviour that he was directly morally responsible for a decision or omission concerning his future behaviour. I think that is extremely implausible. It's overwhelmingly likely that Jeff has reflected on his insensitive behaviour at some point since he was 15 years old. Such moments of reflection may not occur as often in the life of a jerk as they do in the lives of more considerate people, but I think it's implausible to suggest that they never occur.[31] I conclude that Vargas's attack on the tracing principle is unsuccessful.

There are, no doubt, other potential challenges to the tracing principle as I have outlined it.[32] I will not pursue them because my arguments do not depend on the viability of tracing. I think the tracing principle is extremely plausible, and it can feature in many philosophical theories – whether compatibilist, libertarian or sceptical about moral responsibility.[33] But I will distinguish traced moral responsibility from direct moral responsibility *without* assuming that the tracing principle is correct. My arguments concern the conditions for direct moral responsibility. For philosophers who reject the tracing principle, *all* moral responsibility is direct moral responsibility. When discussing particular cases, I will often consider the possibility that agents bear traced moral responsibility for their actions. But I won't need to assume the validity of the tracing principle.

To avoid cumbersome phrasing, I'll often use the abbreviation 'TMR' to stand for traced moral responsibility, and 'DMR' to mean direct moral responsibility. I'll also use those abbreviations as adjectives, so that for example "Trisha is TMR for her action" means that Trisha bears traced moral responsibility for her action.

1.3. OVERVIEW OF THE ARGUMENTS

I'll begin, in chapter 2, by addressing Frankfurt's identificationist theory of moral responsibility. After sketching the development of his account, I'll discuss some of the most important criticisms that it faces, in section 2.3. These criticisms support the conclusion that, contra Frankfurt, his identification condition is *not sufficient* for moral responsibility for action.

I'll set aside those criticisms in chapter 3, in order to build an independent argument against Frankfurt's position, which I call my "main argument". In the first three sections of the chapter, I'll define "subverting causes" and introduce five examples of actions with subverting causes, together with the experimental evidence which supports those examples. (I'll dub these examples "the five key cases", since I will refer to them often.) Then in section 3.4 I'll set out my argument that Frankfurt's identification condition is not sufficient for direct moral responsibility. In the five key cases of actions with subverting causes, the agents meet Frankfurt's identification condition, and yet are not directly morally responsible.

[31] If we found that Jeff never did reflect on his behaviour and its effects, I think we might doubt his normative competence, in the sense I'll discuss in section 2.3.1. We would doubt whether his ethical sensitivity is sufficiently well developed for him to be held morally responsible for any actions at all.

[32] One interesting question, rarely discussed as far as I know, is whether responsibility for a current action can be traced back to *more than one* earlier action or omission.

[33] Vargas comments: "One of the nice features about tracing is that it is one of the few things to which nearly all parties in the debate about free will appeal with equal enthusiasm" (2005:270).

I'll make a "positive proposal" of an alternative necessary condition of direct moral responsibility in section 3.5, and argue that it should be acceptable to Frankfurt and his followers.

In the remainder of chapter 3, I'll discuss experimental evidence which suggests that there are many everyday actions with subverting causes. I'll also review evidence which helps to explain why we often fail to notice when actions have subverting causes, and so we are unaware of how common such causes are.

In chapter 4 I'll discuss and reject some objections to my main argument against Frankfurt and to my positive proposal.

In chapter 5 I'll discuss two philosophers who have engaged with very similar subject matter. Both Nahmias and Doris have proposed identificationist conditions of direct moral responsibility which take account of research in situationist social psychology.[34] I'll argue that their conditions are neither necessary nor sufficient for direct moral responsibility.

I'll move on in chapter 6 to consider Fischer and Ravizza's theory of moral responsibility. I'll explain their conditions for "guidance control" over actions, and then (in section 6.6) discuss some criticisms which have been made of their account. My own argument against their position begins in section 6.7, using actions with subverting causes as counterexamples to show that the guidance control conditions are not sufficient for direct moral responsibility. In the five key cases, the agents meet the guidance control conditions and yet are not directly morally responsible for their actions.

In chapter 7 I make a "sceptical argument": I'll argue that there are many everyday actions for which we mistakenly hold agents morally responsible. Drawing on evidence from chapter 3, I'll argue that there are many everyday actions which have subverting causes. Many of those are actions for which we currently hold agents morally responsible. But, in many of those same actions, the agents are not in fact morally responsible – they bear neither direct nor traced moral responsibility.

Finally, in chapter 8, I'll summarise my arguments and conclusions.

[34] Situationist causes can be subverting causes; but there are also other kinds of subverting cause which have been revealed by research in other fields of psychology.

2. FRANKFURT'S IDENTIFICATIONIST THEORY

One way to characterise Frankfurt's enormous contribution to the problems of freedom and moral responsibility is as a revolutionary departure from the prevailing compatibilist view, often labelled "classical compatibilism". Frankfurt challenged classical compatibilism with two new theoretical proposals, both of which have been extremely influential.[1]

The first of these proposals, which challenges incompatibilists as well as classical compatibilists, is that alternate possibilities are not necessary for moral responsibility. Frankfurt argues that an agent can be morally responsible for his action even if he could not have done otherwise.[2] I won't discuss this argument here. In fact, I will remain neutral on the question whether alternate possibilities are required for moral responsibility: none of my arguments depend on the answer to that question.

I will focus on the second of Frankfurt's theoretical proposals, in which he gives a positive account of free will and moral responsibility. In classical compatibilism, freedom and moral responsibility require an absence of impediments or constraints upon action. In the absence of constraints, the agent is held to be free to act as he wants. In early compatibilist theories, the most commonly cited constraints preventing free action were *physical* constraints such as paralysed legs or locked doors. As understanding grew of psychological conditions such as compulsions and addictions, these were added to the list of constraints which could prevent a free action.[3] A criticism often made of the classical tradition is that this list of constraints is merely ad hoc: the constraints which appear on the list are those which we generally take as preventing free action, but we are not provided with an independent explanation of why certain constraints prevent free action, while others do not.[4]

Frankfurt's theory provides a principled account of *why* certain psychological conditions can prevent freedom and responsibility.[5] Frankfurt distinguishes between *internal* and *external* desires. The agent "identifies himself" or "is identified" with desires which are internal – which are "truly his own". In contrast, an *external* desire constrains the agent and prevents moral responsibility.[6] A drug addict, for example, may feel that "the force [i.e. the desire] moving him to take the drug is *a force other than his own*" – and so is external to him in this sense.[7]

As I'll explain in section 2.1, Frankfurt has made more than one attempt to define exactly what identification consists in. I'll examine the relationship between identification and moral responsibility, in Frankfurt's theory, in section 2.2. Then in section 2.3 I'll examine three important criticisms of his account. The criticism which forms my main argument will be the subject of chapter 3.

2.1. FRANKFURT'S THEORY IN DETAIL

In Frankfurt's own words, the notion of identification is "admittedly a bit mystifying".[8] He has developed his account of identification over a period of three decades, with a series of

[1] I am of course greatly oversimplifying Frankfurt's contribution in this summary.
[2] Frankfurt 1969.
[3] For example Ayer views kleptomania as a constraint on the agent's "process of deciding" which action to perform (1954:20-22).
[4] This is not, of course, the only criticism of classical compatibilism; for overviews of other important issues, see for example Berofsky 2002 and McKenna 2004a.
[5] It's often said that Frankfurt provides an account of *free will*, while classical compatibilism merely gives an account of *free action*.
[6] Frankfurt sometimes describes an external desire as "alien" or "an outlaw".
[7] Frankfurt 1971:18 (my italics).
[8] Frankfurt 1975:54.

articles in which he has gradually changed his position.[9] Unfortunately he has not given a definitive statement of his final position on all the elements of the theory. In particular, while Frankfurt introduced the notion of identification primarily as a condition of freedom and moral responsibility, in later work he has focused increasingly on its role as a condition of psychological well-being.[10]

I'm not certain whether Frankfurt endorses the "tracing principle", by which an agent's moral responsibility for an action can sometimes be traced back to facts about an *earlier* action.[11] Frankfurt's identificationist theory is a theory of *direct* moral responsibility (DMR). If Frankfurt rejects the tracing principle, then he will view *all* moral responsibility as direct moral responsibility. If he accepts the tracing principle, then Frankfurt will allow an agent can be morally responsible for an action when the identification conditions are *not* met, because the agent bears traced moral responsibility (TMR). I won't make any assumption about Frankfurt's stance on tracing. The tracing principle will not be relevant in this chapter; in later chapters I will point out some implications of accepting or rejecting it.

With these caveats in mind, the position I set out in this section is the one which I take Frankfurt to hold. It's a testament to the importance of Frankfurt's proposal that the theory has remained, in its changing forms, at the forefront of critical discussion over such a long period. I'll present Frankfurt's theory as it has developed through three key stages, which I'll call his *initial*, *intermediate* and *final* accounts. It may well be that there have been more than three such stages in the theory's development, but I think the three I describe are the most important. I'll begin by setting out the main elements of Frankfurt's initial account of identification.

2.1.1. Frankfurt's initial account

I'll begin by defining the key terms in Frankfurt's initial account, and then give some examples which will help to explain those terms.[12] A "first-order desire" is a desire to perform an action. An *effective* first-order desire is one which "moves" the agent "all the way to action".[13] Many first-order desires are not effective – usually because they conflict with other first-order desires. A "second-order" desire has as its object a first-order desire. There may be third-order desires which have second-order desires as their objects, and so on. The phrase "higher-order desire" refers to any desire above the first order in this hierarchy. A higher-order *volition* is a desire that a certain first-order desire will be effective. The agent is (directly) morally responsible for an action when there is an alignment or "mesh" between his higher-order volition and his effective first-order desire.[14] In Frankfurt's phrasing, the agent "identifies himself" with that first-order desire, by having the higher-order volition. These definitions are best explained with some examples.

Dennis is in a restaurant studying the dessert menu. He has a first-order desire *G* to order chocolate gateau. He also has another first-order desire *D* to decline the dessert and

[9] Perhaps a more charitable interpretation would be that, over the course of the three decades, Frankfurt has merely *clarified* his initial position, which has remained unchanged. But most commentators portray him as changing his position over a series of papers.

[10] Scanlon makes this point (2002:167).

[11] I discussed the tracing principle in section 1.2. I'm not aware of any passages in which Frankfurt discusses tracing. Fischer and Ravizza seem to interpret Frankfurt as rejecting the tracing principle (1998:194-201). On the other hand, Watson makes it clear that identificationist theories can incorporate tracing (2001:304-305).

[12] Frankfurt's initial account is found in his paper 'Freedom of the Will and the Concept of a Person' (1971). To avoid possible confusion, I have deliberately avoided using the term "will", and the phrases "free will" and "freedom of the will", because Frankfurt employs them in a quite specific way which differs from many other philosophers' usage.

[13] Frankfurt 1971:14.

[14] The useful term "mesh" was introduced by Fischer (1987:79).

head home instead. He is on a diet, and has a second-order volition that desire *D* be effective. Desire *D* moves him to action, and he declines the dessert. His second-order volition meshes with the effective first-order desire: he is morally responsible for the action.

Ulf is an "unwilling addict" about to inject his usual drug.[15] He has a first-order desire *T* to throw away the needle, and at the same moment a first-order desire *F* to inject his next fix. His second-order volition is that desire *T* be effective. Instead, desire *F* is effective – he injects the fix. There is no mesh between second-order volition and effective first-order desire: Ulf is not morally responsible for this action.

William is a willing addict. He has the same first-order desires *T* and *F* as Ulf. William's second-order volition is that desire *F* be effective, and it is – he injects the fix. The mesh obtains: William is morally responsible for his action.[16]

Being capable of second-order volitions is essential to being a *person*. Very young children and animals have no second-order volitions, and so are termed "wantons". Adults too may *act wantonly* in response to first-order desires about which they have no second-order volitions. A "wanton addict" has a first-order desire to inject a fix; he may also have a first-order desire to refrain from taking it. But he has no higher-order volition which favours either one of these first-order desires. Whichever first-order desire is effective here, "he has no stake in the conflict",[17] and so he does not act as a morally responsible person.

There is "no theoretical limit" to the number of levels in Frankfurt's hierarchy of desires. Sometimes, for example, an agent may have a third-order desire about one of his second-order desires. Frankfurt recognised the threat of a damaging regress. Why does a mesh between an agent's first-order desire and his second-order volition sometimes suffice for identification? Why isn't a mesh with third-order volitions always necessary? And if third-order desires are sometimes necessary, why not even higher orders of desire … and so on?

Frankfurt initially claimed that the regress could be terminated without arbitrariness:

> When a person identifies himself *decisively* with one of his first-order desires, this commitment "resounds" throughout the potentially endless array of higher orders … there is no room for questions concerning the pertinence of desires or volitions of higher orders.[18]

But, at least in Frankfurt's initial presentation of his theory, it is unclear how a resounding "commitment" would prevent the damaging regress. Suppose that William the willing addict wants (second-order desire) his first-order desire *F* to inject a fix to be effective. On Frankfurt's account, such an arrangement can constitute a decisive identification. But now consider Wanda, who also has a second-order desire that her first-order desire *F* to inject a fix will be effective. Wanda's second-order desire is itself brought on by her addiction, and she has no third-order desire about this second-order desire. It would seem quite plausible to say that Wanda's second-order desire is *wantonly* held. And if the second-order desire is wantonly held, then there is no decisive identification.

Why does William's case involve genuine identification, while Wanda's does not? Why is the lack of a third-order desire significant in Wanda's case, but not in William's? Frankfurt's initial account lacks the resources to answer this question. As Gary Watson argued:

[15] The "unwilling addict" is a key example in Frankfurt's 1971 argument.
[16] See Frankfurt 1971:25.
[17] Frankfurt 1971:19.
[18] Frankfurt 1971:21.

> It is unhelpful to answer that one makes a "decisive commitment," where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary.[19]

It seems that Frankfurt later accepted this point,[20] and made amendments over a series of papers to take up what I will call his *intermediate* position.[21]

2.1.2. Frankfurt's intermediate position

By 1987, Frankfurt's answer to the regress problem involved *decisions* about desires:

> it is characteristically by a decision … that a sequence of desires or preferences of increasingly higher orders is terminated.[22]

Once the agent has taken a decision in favour of some higher-order desire, there is no longer an open question about whether that desire would be endorsed by a desire of a still higher level. What *is* a decision about a desire? Frankfurt describes a commitment made "without reservation", such that

> the person who makes it does so in the belief that no further accurate enquiry would require him to change his mind. It is therefore pointless to pursue the enquiry any further.[23]

By making such a decisive commitment to a higher-order desire, the agent identifies with the first-order desire at the bottom rung of that hierarchy.

> The decision determines what the person really wants by making the desire on which he decides fully his own.[24]

Frankfurt claims that a decision in favour of a desire would eliminate the threat of an infinite regress of desires, which was a problem for his first account.[25] The basis for this claim is not entirely clear to me. There seem to be two possibilities. First, Frankfurt may be claiming that a *decision* in favour of a desire is not itself capable of being the object of a desire. There would then be no possibility of regress beyond the decision, since there could be no higher order of desires above the decision. But it seems false to say that decisions (even decisions about desires) cannot be the objects of desires. It's surely possible – indeed common – for us to have desires about what decisions to make. If this is correct, then a decision which is made in the absence of desires about the decision appears to be wanton. And wantonness precludes identification.

The second possibility is that Frankfurt believes that a decision in favour of a desire must be internal to the agent, and cannot be external.[26] If that is correct, identification would be entailed in the moment of deciding. However, it seems that decisions about desires *can* be

---

[19] Watson 1975:29. In the same article Watson proposed his own alternative account of identification. Very roughly, on this account an agent identifies with an effective desire to perform an action if and only if that action is what she most values performing. However, Watson's account succumbed to the problem that an agent can identify with an action even when it is not what she values performing, as he later acknowledged (1987b).

[20] See Frankfurt 1977:65-66.

[21] The most important developments between the initial account and the intermediate position are described in Frankfurt's 1975, 1977, and 1987a papers.

[22] Frankfurt 1987a:170.

[23] Frankfurt 1987a:168-169.

[24] Frankfurt 1987a:170.

[25] Frankfurt denies that Watson's charge of arbitrariness applies to decisions about desires (1987a:167).

[26] Frankfurt had earlier claimed that "Decisions, unlike desires or attitudes, do not seem to be susceptible both to internality and to externality" (1977:68n3).

external to the agent. Frankfurt cannot rule out, as far as I can see, that a decision about a desire might be made as a result of an agent's *compulsion*, and so not truly her own.

Furthermore, a decision in favour of a first-order desire might be "unwitting", and so not truly the agent's own. David Velleman describes an example in which he meets an old friend, aiming to resolve a minor disagreement. He finds himself annoyed by the friend's attitude and begins to raise his voice, and the two men part in anger. On later reflection he realises that "accumulated grievances had crystallized … into a resolution to sever our friendship over the matter at hand".[27] There is nothing abnormal or compulsive about the desires in this case. Yet Velleman believes that the resolution is not truly his own: "it was my resentment speaking, not I".[28] Velleman concludes:

> the example of my unwitting decision to break off a friendship shows that even decisions and commitments can be foreign to the person in whom they arise.[29]

Thus, if Velleman is right, identification is not entailed even by a (non-compulsive) decision which meshes with an effective desire.

In response to this problem case, it might seem open to Frankfurt to *define* the crucial notion of a "decisive commitment" as one which the agent necessarily regards as internal to him. But this response would be circular: identification would require a kind of decision which itself requires identification.[30]

A second possible response for Frankfurt would be to claim that the effective first-order desire in Velleman's example is *not* in fact supported by a *decision*. Instead, he might claim, the resolution results from some mental event which is something less than a decision – perhaps we might call it an "opting" event.[31] This response would require that there is some distinction between decision and opting which is independent of the notion of identification (in order to avoid the circularity in the first response). In this vein one might claim that "decisions" about desires must involve careful conscious deliberation, of a kind which will ensure that any effective desires endorsed by decisions are internal. That may be a plausible definition of "decisions". But this line of thinking seems to lead toward scepticism about the prevalence of moral responsibility. For there are surely many actions which we perform without such careful conscious deliberation about our desires; compatibilists will reject a definition of identification which suggests that agents are not identified with their desires in a lot of everyday actions.

It may be that Frankfurt was influenced by Velleman's paper.[32] At any rate, he later accepted that no particular "deliberate psychic element" can be essential to identification.[33] *Any* such deliberate psychic element – including a decision – would be one about which the agent could have a conflicting higher order desire or attitude. And so a problematic regress threatens once more. In response, Frankfurt moved to what I'll call his *final* account of identification.

---

[27] Velleman 1992:192.
[28] Velleman 1992:192.
[29] Velleman 1992:200. The word "unwitting" may perhaps mislead here. The point is not that the agent lacked *awareness* of the decision or the desire; but rather that the effective desire was not truly his own, in spite of the decision in its favour.
[30] Velleman makes this point (1992:200n27).
[31] I've avoided the term "choice" here, which is used by Frankfurt in a particular technical sense (1987a:172).
[32] Bratman (1996:193) draws a link between Velleman's paper and Frankfurt's abandonment of the intermediate position.
[33] See Frankfurt 1992:104. "Psychic elements" include beliefs, desires, attitudes, judgements and decisions.

2.1.3. Frankfurt's final account

The final account is by far the most complex and subtle.[34] Frankfurt introduces two new elements into the theory of identification – "acceptance" and "satisfaction". In the initial account, identification with an effective desire consisted simply in having a higher-order desire which meshes with the effective desire. But that account succumbed to the threat of an endless regress of higher orders of desires. Frankfurt now builds on that initial account by adding an extra condition, which introduces the element of satisfaction.

> The endorsing higher-order desire must be, in addition, a desire with which the person is *satisfied*.[35]

Being satisfied with one's desires is "a matter of simply *having no interest* in making changes".[36]

How is the regress avoided, on this account? There is no requirement that the agent have any psychic element which meshes with the endorsing higher-order desire. Identification in fact depends on the *absence* of any psychic element which *conflicts* with that desire. Thus,

> being satisfied with a certain desire does not entail an endless proliferation of higher order desires. Identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied.[37]

Now it may be that there *is* some specific judgement (or other psychic element) which meshes with the endorsing higher-order desire. But even the presence of such a judgement does not ensure that the agent is satisfied. For example, judging it best not to try to change one's desires is not the same as being genuinely satisfied with them.[38]

At the same time, it's possible to be satisfied with one's desires (even first-order desires) "without in any way considering whether to endorse them".[39] But the absence of interest in making a change must be "reflective" – it "must derive from [the agent's] understanding and evaluation of how things are with him".[40] A deliberately contrived or wanton lack of interest in making changes does not constitute genuine satisfaction.

Frankfurt's use of the term "endorse" is somewhat misleading, as he later acknowledges.[41] He does not mean that the endorsed desire must be positively evaluated, or approved of. The sense of endorsement he has in mind requires "nothing more than that the agent *accepts* it as his own".[42] Such acceptance is an "altogether neutral attitude":

> A person may be led to accept something about himself in resignation, as well as in approval or in recognition of its merit. The fact that he accepts it entails nothing, in other words, concerning what he thinks of it.[43]

This is an important change to the theory. In the initial and intermediate accounts, the agent's top-level desire or decision included a *normative* assessment of the lower level desires. But

---

[34] The exegetical problems are especially severe in the final account, which was first introduced in Frankfurt's 1992 paper. In a collection of essays on Frankfurt's work, it's clear that two extremely eminent commentators – Bratman (2002a) and Watson (2002) – misinterpreted important elements of his (final) theory of identification. See Frankfurt 2002b and 2002c for his clarifications. I have drawn heavily from these papers, as well as Bratman 1996, Velleman 2002 and Taylor 2005, for my summary here.

[35] Frankfurt 1992:105.

[36] Frankfurt 1992:105.

[37] Frankfurt 1992:105.

[38] Frankfurt 1992:104.

[39] Frankfurt 1992:105.

[40] Frankfurt 1992:105.

[41] See Frankfurt 2002b:87 and 2002c:160.

[42] Frankfurt 2002b:87.

[43] Frankfurt 2002c:160.

the "neutral attitude" of acceptance simply marks the agent's *descriptive* assessment about whether the lower-order desires are his own.[44]

To summarise, an agent identifies with an effective desire to act if he *accepts* the desire as his own, and if he is *satisfied* with that attitude of acceptance. It is possible for an agent to exhibit acceptance without satisfaction. This can happen in a condition which Frankfurt calls "ambivalence". Here two desires or volitions are inherently opposed, in the sense that they cannot both be fulfilled; yet both are accepted as the agent's own. For example, someone may want to commit to a certain career, but also want to refrain from doing so.[45] According to Frankfurt, this person cannot be satisfied with his "psychic condition".

In contrast to ambivalence, "wholeheartedness" consists in being satisfied that certain desires and volitions should be effective, while others are not.[46] Wholeheartedness is compatible with conflict among one's desires. For example, an unwilling addict may be wholehearted if he is satisfied in accepting one set of desires and volitions as his own, while rejecting conflicting desires to take drugs as external.[47] But a wholehearted agent does not *accept* two or more conflicting desires as his own.

For simplicity moving forward, I'll introduce a technical term "F-accepts", whose meaning is as follows. An agent "F-accepts" his effective desire if he *accepts it as his own*, and *is satisfied with that acceptance*, where being satisfied is a matter of having no interest in making changes. According to Frankfurt's final account, an agent identifies with his effective desire if and only if he F-accepts it.[48]

One might object that Frankfurt has failed to avert the vicious regress which overwhelmed his earlier accounts. It's not obvious that "reflective" satisfaction can be distinguished clearly from a *merely wanton* lack of interest in making changes. The agent must have reached an "understanding and evaluation of how things are with him". But he must not have formed a *desire* not to change his desires, nor a *judgement* that change is not appropriate: the presence of either of these would open up the possibility of regress. What kind of reflective process can the agent undergo which would produce this outcome? Perhaps Frankfurt has in mind a kind of process which *might* produce a higher-order psychic element, but in fact does not do so. It's not obvious that such a fine balance is possible, let alone common enough to enable agents to be identified with most of their actions. However, I'll set aside this objection and assume, for the sake of my later arguments, that Frankfurt succeeds in avoiding regress in his final theory.

2.1.4. Terminological simplifications

Strictly, on Frankfurt's final account, the object of the agent's F-acceptance is *the effective desire to act*. This is also what the agent is identified with. Now, the first-order desire which is effective has as its object *the action performed*.[49] There is a one-to-one relationship between the effective desire to act and the action performed.

Therefore, the phrase "William F-accepts the effective desire which moves him to act" is equivalent, for practical purposes, to the phrase "William F-accepts his action". For reasons

[44] Taylor makes this point (2005:9).
[45] The example is Frankfurt's (1992:99).
[46] Frankfurt 1992:103.
[47] Frankfurt 1992:100.
[48] Frankfurt says that identification "consists in" having an "endorsing" (by which he means "accepting", as he later clarifies) higher order desire with which the agent is satisfied (1992:105).
[49] The desire *is effective* when it moves the agent to perform the action desired.

of simplicity and style, I'll often use the shorter phrasing in place of the longer version. Similarly, I will often use the phrase "William identifies with his action" rather than the more precise but more cumbersome phrase "William identifies with the effective desire which moves him to act". I'll also sometimes say that "William's action is truly his own", rather than "the effective desire which moves William to act is truly his own".

I don't think I risk any misunderstanding of my argument by making these simplifications.

2.2. IDENTIFICATION AND MORAL RESPONSIBILITY

My exegetical caveats extend to Frankfurt's (final) view of the relationship between identification and moral responsibility. There are four pieces to this puzzle, and not all the logical relations between them are explicitly described by Frankfurt, it seems to me. In their positive forms, I'm interested in the following four states of affairs:

(A) The agent *F-accepts* his effective desire.[50]
(I) The agent *is identified with* his effective desire.
(O) The agent's effective desire *is truly his own*.[51]
(R) The agent is *directly morally responsible* for the action to which he is moved by his effective desire.

I'll assemble a set of six statements which I'm confident that Frankfurt holds, and which jointly define Frankfurt's view of the logical relationships between these states of affairs.[52] I'll list the full set of statements at the end of this section.

In Frankfurt's final position, (A) is necessary and sufficient for (I): an agent *is identified with* his effective desire if and only if *he F-accepts it*. I'll label this statement (1) in the set which defines Frankfurt's position.

I'm rather less confident of Frankfurt's full view of the relations between (I) and (O), and between (I) and (R).[53] Fortunately, he does make two clear declarations which establish at least part of the picture. Both come in replies to critics of his position, in which he also clarifies some important elements of the final theory, so it seems reasonable to take them as definitive.

Frankfurt says:

> A desire with which an agent identifies is legitimately attributable to him as his own; it is not external to him or, in other words, an outlaw.[54]

From this I conclude that (I) is sufficient for (O): if the agent *is identified with* his effective desire, then that desire *is truly his own*. This is statement (2) in my set.

Frankfurt also states that:

> Someone who is wholeheartedly behind the desires that move him when he acts is morally responsible for what he does.[55]

---

[50] An agent F-accepts his effective desire if he accepts it as his own, and is satisfied with that acceptance. Being satisfied is a matter of having no interest in making changes. See section 2.1.3.
[51] When a desire is "truly his own" it is *internal* to the agent, rather than *external* or "alien".
[52] Occasionally I will use a logical notation to condense a brief argument. In that notation, the symbol "→" means "entails that", or "is sufficient for"; "↔" means "if and only if" or "is necessary and sufficient for"; and "¬" means "it is not true that".
[53] As I'll discuss below, I am not certain whether Frankfurt regards (I) as necessary for either (O) or (R).
[54] Frankfurt 2002b:88.
[55] Frankfurt 2002d:28.

From this I conclude that (I) is sufficient for (R): if the agent *is identified with* his effective desire, then he is *directly morally responsible* for the action to which he is moved by that effective desire. This is statement (3) in my set.

It is not clear to me whether (I) is necessary for either (O) or (R), in Frankfurt's final position. As Teresa Chandler points out, Frankfurt's view of the relationship between (I) and (O) appears to have changed over time.[56] In Chandler's phrasing, Frankfurt moves from an "invention account" to a "discovery account" of identification. In the early "invention account", identifying with the desire is what makes the desire truly one's own. This is especially clear in what I called Frankfurt's intermediate position, in which identification involves a decision which "determines what the person really wants".[57] Here it seems that (I) is necessary for (O): the desire is truly the agent's own only if he is identified with it.

However, in the later "discovery account" (which corresponds with what I've called Frankfurt's final position), it seems that an agent can have desires which are truly her own, but are not yet F-accepted by her. She may struggle against them, not realising that they are truly her own or refusing to accept them as such, before finally recognising them, and F-accepting them. Frankfurt describes a mother who thinks it best to give up her child to adoption, and decides to do so, but when the moment arrives she finds that she cannot go through with it.[58] The mother discovers something about herself, and specifically about the structure of her own higher-order desires. She may finally F-accept the desire to keep the child. It seems that the desire to keep the child was truly her own, even when the mother was struggling against it and did not F-accept it. It seems that (A) is not necessary for (O), and hence that (I) is not necessary for (O).[59]

Similarly, it may be that identification is not necessary for responsibility – i.e. (I) is not necessary for (R) – on Frankfurt's final view. I'm not aware of any statement or denial by Frankfurt on this point. Therefore I will not assume that (I) is necessary for either (O) or (R), on Frankfurt's view.

Many commentators interpret Frankfurt as holding that (O) is both necessary and sufficient for (R). For example Sarah Buss and Lee Overton, introducing a collection of essays about Frankfurt's work, say:

> According to Frankfurt, we are responsible for what we do if and only if our motives for doing it are truly our own.[60]

But I have not found a clear statement to that effect in Frankfurt's own writing. In particular, it may be that (O) is not sufficient for (R) in Frankfurt's final account. (If Frankfurt held that (I) is necessary for (O), as he seemed to on the "invention" account of identification, then it would follow that (O) is sufficient for (R).[61] But I am not convinced that this is the case in Frankfurt's final account, where identification is more a matter of discovery than invention.)

I will not assume that Frankfurt holds that (O) is sufficient for (R). Instead, I will simply take it that he regards (O) as *necessary* for (R): in other words, that an agent is directly

[56] Chandler 2004:38-43
[57] See section 2.1.2.
[58] The example is originally mentioned in Frankfurt 1993:111, but is developed further by Watson (2002:147) and in Frankfurt's reply (2002c:161-164).
[59] If ¬(O→A) and (I↔A), then ¬(O→I).
[60] Buss and Overton 2002:xii.
[61] We know, from statement (3), that (I→R). If also (O→I), then it would follow that (O→R).

morally responsible for her action *only if* her effective desire is truly her own.[62] This is statement (4).

Since (A) is sufficient for (I), and (I) is sufficient for both (O) and (R), it follows that (A) is sufficient for both (O) and (R). So if the agent *F-accepts* his effective desire, then that that desire *is truly his own*. This is statement (5) in my set. Furthermore, if the agent *F-accepts* his effective desire, then he is *directly morally responsible* for the action to which he is moved by that effective desire. This is statement (6).

To summarise, the six statements which I'm confident that Frankfurt holds are:[63]

(1) F-acceptance is necessary and sufficient for identification (A↔I).
(2) Identification with an effective desire is sufficient for its being truly the agent's own (I→O).
(3) Identification is sufficient for direct moral responsibility (I→R)
(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility for the action (A→R).

2.3. CRITICISMS OF FRANKFURT'S ACCOUNT OF IDENTIFICATION

My main objectives, in arguing against Frankfurt, are to show that statements (5) and (6) above are false.

In my "main argument", which I'll make in chapter 3, I'll aim to show first that statement (5) is false. From that, it follows that (6) is false, if (4) is true.[64] I will only reject three of the six statements that Frankfurt holds: (5), (6), and also (1) – which is Frankfurt's definition of identification. I won't deny (2) or (4), which are both key axioms of identificationist accounts of moral responsibility. For the sake of my main argument, I won't deny (3) either (although I think it is false).

Meanwhile, in the rest of this current chapter, I'll briefly discuss three criticisms which *also* suggest, in different ways, that (5) and (6) are false. The first two criticisms – in sections 2.3.1 and 2.3.2 – aim to show that (3) is false. From that, it follows that (6) is false, given Frankfurt's own definition of identification (1).[65] The third criticism – in section 2.3.3 – denies statement (5). I won't do much more than give an overview of these criticisms, since my main argument does not rely on their conclusions. By including them I hope to achieve two things. First, I'll show that the conclusions of my main argument can be reached in other ways, which gives independent grounds to think that they are, at least, plausible conclusions. Second, discussion of these criticisms introduces some issues which will be relevant in later chapters.

---

[62] It's clear from Frankfurt's discussion of unwilling addicts, for example, that he regards (O) as necessary for (R). An unwilling addict is not morally responsible for her action of injecting a fix *because* the effective desire is not truly her own.

[63] I've abbreviated these statements to avoid cumbersome phrasing, and to incorporate the simplifications I discussed in section 2.1.4. The full version of statement (1), for example, is "the agent's F-acceptance of the effective desire is necessary and sufficient for his identification with the effective desire". I don't think any confusion will arise from my abbreviations.

[64] If ¬(A→O) while (R→O), then it follows that ¬(A→R).

[65] If (A↔I), but ¬(I→R), then it follows that ¬(A→R).

2.3.1. Normative competence

In a very influential paper, Susan Wolf criticised identificationist accounts of responsibility, including Frankfurt's.[66] One of her examples involves JoJo, the son of an evil and sadistic dictator. As a child, JoJo witnesses his father's appalling actions every day. Later, as an adult, he behaves in similar ways. For example, he sends a person to be tortured on a whim. JoJo is not constrained or coerced in this action. Moreover, the effective desire which moves him to the action is one with which he wholly identifies, for he has been raised by his father to develop higher-order desires which endorse such actions.[67]

Wolf concludes that JoJo is not morally responsible for his action, despite being identified with the effective desire. She notes the legal principle, known as the "M'Naughten Rule", that a person is sane only if he knows what he is doing and whether it is right or wrong.[68] JoJo fails to meet the condition of "sanity":

> the minimally sufficient ability cognitively and normatively to recognize and appreciate the world as it is.[69]

In her later book, Wolf substitutes the phrase "normative competence" – with a very similar definition – for this ability.[70]

I think this simple argument is convincing: normative competence is indeed necessary for direct moral responsibility.

2.3.2. The agent's history

Perhaps the most common criticism of Frankfurt's theory is that his identification condition does not take the agent's history into consideration. Consider the following example. Brian was abducted several years ago by some members of a cult and prevented from leaving their remote community. Since that time, the cult leader has used various psychological manipulation techniques (sometimes called "brainwashing") to influence Brian's desires. Brian no longer wants to leave the cult. In fact, Brian writes a new will in which he leaves his estate to the cult leader, rather than to his own relations. Brian is identified with his effective first-order desire, on any of Frankfurt's accounts of identification. And yet a strong intuition tells us that Brian is not morally responsible for this action, because of his past conditioning. The higher-order desire which meshes with the effective first-order desire seems not to be genuinely Brian's own desire, in a very important sense.

This is an issue which divides compatibilists.[71] Some, often called "soft compatibilists", believe that a condition about the agent's history is necessary for moral responsibility.[72] The following are examples of histories which many soft compatibilists (and many incompatibilists) believe can prevent moral responsibility for action: psychological manipulation by a cult leader;[73] behavioural conditioning;[74] an extremely deprived

---

[66] Wolf 1987.

[67] Wolf addresses her criticism to Frankfurt's initial account of identification, but it could be updated to make the same points against his final account.

[68] Wolf 1987:381.

[69] Wolf 1987:381. Wolf notes that this definition of sanity is specifically intended to address questions about *responsibility*, and may not be applicable for use in other contexts.

[70] Wolf 1990:129.

[71] Watson presents the issue in the form of a dilemma for compatibilists: advocates of both positions face difficult problems (1999:209-215).

[72] The terms "hard" and "soft" compatibilists were introduced by Kane (1996:67-68).

[73] Kane argues that agents who have undergone this kind of "covert nonconstraining control" lack autonomy (1996:64-65).

upbringing;[75] hypnosis designed to induce a certain desire or belief.[76] I'll use the generic term "conditioning" to refer to these various kinds of insidious influences. The challenge faced by soft compatibilists is to explain how conditioning in the agent's past might prevent moral responsibility, while the presence of causal determinism throughout the agent's past does not.[77]

Frankfurt, on the other hand, belongs in the "hard compatibilist" camp. Hard compatibilists propose that it is sufficient for moral responsibility that certain conditions are fulfilled *at the time of acting*. They face the problem that we feel a very strong intuition that some events in an agent's history – such as Brian's conditioning in the cult – can excuse him from moral responsibility. Frankfurt is of course well aware of this difficulty for his theory, and his response is staunchly hard compatibilist.[78]

> A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. We are inevitably fashioned and sustained, after all, by circumstances over which we have no control. The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents.[79]

While I admire Frankfurt for taking a principled stance on this point, he seems to be simply ignoring strongly held contrary intuitions. A more persuasive way to deal with intuitions which conflict with one's theory is to acknowledge their *validity*, while explaining why they are *not applicable* in the situations covered by the theory. Frankfurt's own extremely influential argument that alternate possibilities are not required for moral responsibility works in just this way.[80] In contrast, Frankfurt's defence of hard compatibilism amounts to an unconvincing flat denial or rejection of our intuitions.

I think these arguments reveal that some condition concerning the agent's history is necessary for direct moral responsibility.

2.3.3. Desires which are not truly one's own

Another problem for Frankfurt's final account is that there seems to be scope for an agent to F-accept a desire which is *not* truly her own.

---

[74] The fictional inhabitants of Walden Two have undergone extreme behavioural conditioning (Skinner 1948/1976).

[75] A persuasive real-life example is the harrowing account of the childhood and later crimes of Robert Alton Harris given in Watson 1987a. A fictional example is that of JoJo, the son of a sadistic dictator, discussed above (Wolf 1987:379-380).

[76] A fictional situation involving hypnosis is discussed by Slote (1980:149).

[77] Pereboom presents a particularly stark version of the problem for soft compatibilists in his "Four Case Argument" (2001:110-120).

[78] He has taken the same principled line since proposing his initial theory – see Frankfurt 1975:52-54.

[79] Frankfurt 2002d:28.

[80] Frankfurt (1969) argues that our intuitions lead us to a mistaken belief in the 'Principle of Alternate Possibilities'. In many cases of constrained action, we assume that the agent is not morally responsible *because* he is unable to do otherwise. But usually, in these cases, the agent cannot act otherwise *and* does not identify with the action. In certain highly unusual "Frankfurt-style" cases, agents who do identify with their actions can be morally responsible (according to Frankfurt) *despite* having no alternate possibilities. In this argument Frankfurt does not deny or ignore an important intuition about constrained actions; he aims to reveal that we apply it mistakenly.

Suppose that Christine is extremely self-critical. She strongly believes in taking the blame for one's own faults. She is a drug addict, and she dislikes her own habit of taking drugs. But she has come to accept that the effective desires and volitions in favour of taking drugs are her own. She is satisfied with that acceptance. And yet it seems that she might be mistaken here: she might F-accept the desires *in error*. Suppose that she discusses her situation with an addiction therapist, and then comes to the opposite opinion that the desires are not her own. We might plausibly conclude, I think, that her new opinion is correct and her initial F-acceptance was mistaken. Her desires and her normative judgements about those desires need not have changed at all, but her descriptive assessment now seems to be more accurate.

The same issue arises when agents simply form different descriptive assessments at different times. For example, Derek is depressed and accepts as his own the desire to quit his profession, even though he doesn't think that quitting is the best thing to do.[81] He thinks to himself "I might as well accept it. I'm just a quitter". He is satisfied with this acceptance, and quits his profession. But after his depression has lifted, he takes a very different view, and thinks "I was not myself when I decided to do that". Frankfurt must hold that, since Derek F-accepted the effective desire at the time of acting, it was truly his own. And yet Derek himself now rejects that conclusion, and people who know him well might take the same view.

It seems possible that agents could form different descriptive assessments of their own desires even at quite short intervals. Someone in a particularly ecstatic or fearful mood might F-accept a desire, only to form a different assessment an hour later. Someone who is drunk or under the influence of drugs might F-accept a desire but later renounce that acceptance. In these cases, too, the later assessment seems to reflect the agent's more "normal" state, in the sense that seems most relevant for judging whether the desire was truly her own.

It might seem that Frankfurt can reply that agents in such situations are not "wholehearted" when they accept their effective desires at the time of acting. But since "wholeheartedness" consists in being satisfied that certain desires and volitions should be effective, I don't think this reply would help: in F-accepting a desire, the agent is *already* satisfied with it. At issue here, as Laura Ekstrom points out, is how to interpret Frankfurt's notion of *satisfaction*.[82] If satisfaction is "a kind of *feeling*" (or perhaps the absence of a kind of feeling), then it seems that Christine and Derek are indeed satisfied with their effective desires. Satisfaction might instead be interpreted as "a *structural* requirement" of "cohesion" among the agent's volitions.[83] But it seems to me that satisfaction is portrayed as a *feeling* on the most natural reading of Frankfurt's 1992 paper 'The Faintest Passion'. At any rate, it's not obvious how Frankfurt could incorporate *cohesion* into the notion of satisfaction, since he allows that satisfaction and wholeheartedness are compatible with conflict among one's desires.[84]

I conclude that, in these cases, F-acceptance of a desire does not ensure that the desire is truly the agent's own, in the deep sense which Frankfurt aims to analyse.

2.3.4. Interim conclusions on Frankfurt's account

In section 2.2 I assembled the following set of statements to summarise Frankfurt's (final) position on identification and responsibility:

[81] This case is adapted from Ekstrom 2005:148. Bratman makes the very similar point that an agent may simply F-accept a desire through *exhaustion*, while the desire is not truly his own (1996:194-195). Frankfurt appears to hold that F-acceptance "in exhaustion" is identification (2002b:90n2).

[82] Ekstrom 2005:143-144.

[83] Such an interpretation would bring Frankfurt's theory closer to Ekstrom's.

[84] I discussed an example above: an unwilling addict may be wholehearted if he is satisfied in accepting one set of desires and volitions as his own, while rejecting conflicting desires to take drugs as external.

(1) F-acceptance is necessary and sufficient for identification (A↔I).
(2) Identification with an effective desire is sufficient for its being truly the agent's own (I→O).
(3) Identification is sufficient for direct moral responsibility (I→R).
(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility (A→R).

I've discussed three criticisms of Frankfurt's theory of identification. The first two show that identification is not sufficient for direct moral responsibility: conditions concerning the agent's history and normative competence are also necessary. I conclude from this that Frankfurt's claim (3) is false. If statement (3) is false, then on Frankfurt's own definition of identification (1), it follows that (6) is false: F-acceptance is *not* sufficient for direct moral responsibility.[85]

The third criticism suggests that F-acceptance of an action is not in fact sufficient for the action's being truly one's own. Actions can be F-accepted even though they are not truly one's own. I conclude from this that claim (5) is false.

These criticisms, then, give support to my main conclusion that both (5) and (6) are false.

However, I haven't argued thoroughly for these conclusions in this chapter, so I will not rely on them as I move forward.

Instead, in the next chapter I'll give an independent argument that Frankfurt's claims (5) and (6) are false. None of the examples I'll use to build that argument will involve agents who have a responsibility-undermining history or lack normative competence, nor agents who are depressed, exhausted or unusually self-critical. Instead, I'll argue that when an action has subverting causes, that action is not truly the agent's own, and so she is not directly morally responsible for it – even when she F-accepts it.

[85] If ¬(I→R), while (A↔I), then ¬(A→R).

3. ACTIONS WITH SUBVERTING CAUSES

In the previous chapter I summarised Frankfurt's theory of direct moral responsibility for actions.[1] An agent is *identified* with his effective desire, on Frankfurt's account, if and only if he accepts it as his own, and is satisfied with that acceptance, where being satisfied is a matter of having no interest in making changes. I introduced the term "F-acceptance" to stand for this combination of acceptance and satisfaction.

According to Frankfurt, F-acceptance of an effective desire is sufficient for direct moral responsibility for the resulting action.[2] I'll argue that this is false. In some actions with what I call "subverting causes", the agent may F-accept her effective desire, and yet not be DMR for the action.

I'll begin by describing five examples of actions with subverting causes (section 3.2). Since I'll refer to those examples many times, I'll dub them "the five key cases". I'll discuss some of the most important features of subverting causes in section 3.3.

I make my main argument against Frankfurt's identification condition in section 3.4. Then, in section 3.5, I make a positive proposal of an alternative condition which I believe to be necessary for direct moral responsibility.

The five key cases provide only a small sample of the evidence that actions do have subverting causes. I'll discuss the larger body of evidence discovered through psychology research in section 3.6, followed by objections to my claims about this evidence in 3.7. Then in section 3.8 I'll review evidence which explains why we often fail to notice that some of our actions have subverting causes.

I'll finish by summarising the chapter's most important conclusions in section 3.9.

3.1. DEFINITION

As a simplified working definition, an action has subverting causes when it is true that: *if the agent knew all the causes of her action, she would not identify with the action*. Since I am arguing against Frankfurt's (final) position, I'll substitute his identification condition into this working definition. An action has subverting causes when it is true that:

> if the agent knew all the causes of her action, she would not F-accept the action.

The definition incorporates what I'll call a "counterfactual test": the verdict hinges on what would be the case if the agent knew all the causes of her action. It's possible that, although in fact the agent *does* F-accept her action, she would no longer do so if she knew all of its causes.

The full definition is as follows. An action has subverting causes when it is true that:

> if, at the time of acting, the agent had known all the causes of her action (except in so far as any additional knowledge gained would have given her new information about means to satisfy her desires), she would not have F-accepted the effective desire which moved her to act.

Compared with the working definition, the full definition includes three additional clauses. The first specifies that the relevant counterfactual test is made *at the time of acting*. The second clause (in parentheses) constrains the type of knowledge that the agent gains in the counterfactual scenario. Briefly, this second clause excludes knowledge which would make

[1] Direct moral responsibility (DMR) is anchored in facts about the action and the agent at the time of acting. In contrast, traced moral responsibility (TMR) can be "traced back" to facts about an earlier action (or omission) and about the agent at that earlier time. See section 1.2.

[2] This is statement (6) from section 2.2.

the agent *regret* her action, yet still regard it as truly her own.[3] I'll discuss this clause in detail in section 4.2. Finally, on the more strictly precise definition, the object of the agent's counterfactual F-acceptance is "the effective desire which moved her to act", rather than "the action" itself.[4] I'll usually avoid this longer and precise but clumsy phrasing, in favour of the simpler but slightly less precise working definition above.

I include among the "causes" of the action any proximate, relevant event *without which the action would not have occurred*. I'll discuss the notions of proximity and relevance, as well as some other important features of subverting causes in section 3.3 below. For simplicity, I'll often use the word "causes" as a shorthand for "proximate and relevant causes".

Strictly speaking, if the causes of an action are subverting, they are subverting *as a set*. However, it's often possible to identify a certain *individual* cause without which the set of causes would not be subverting. When that is the case, for simplicity of style, I'll call that individual cause "the subverting cause".

3.2. SUBVERTING CAUSES – FIVE KEY CASES

I'll now present five key cases of actions with subverting causes, which will feature in my main argument and much of the subsequent discussion. As I'll point out, each case closely matches an actual research experiment. The first three cases are taken from the field of social psychology known as "situationism".

3.2.1. Example 1: Situational factors and bystander intervention

Samantha is on her way to give a short presentation about the parable of 'The Good Samaritan'.[5] She heads down an alley and walks past a man who is slumped in a doorway apparently in distress. A cause of her action is:

(a) A few moments ago Samantha was told that she was running late for the presentation.

If she had not been told that she was running late, Samantha would have stopped to help the person slumped in the doorway.

This example matches a study by psychologists Darley and Batson (1973), whose subjects were students at a theological seminary. Some were told that they were a little early for their presentation: 63% of this group stopped to offer help to the slumped man (who was a confederate or 'stooge' of the experimenters). Of those subjects told that they were running late, only 10% stopped. Some of the subjects who walked past the slumped man appeared anxious about their decision, while others appeared to assess him as not in need of assistance.[6]

Telling the subjects that they were early or late was a "situational" factor which varied. The experimenters also varied another situational factor – whether the subjects were to give the presentation about the parable of the 'Good Samaritan', or whether they were to speak instead about job prospects for seminarians. The experimenters also tested the subjects on a personality measure, via a questionnaire. Only the first factor – the early or late condition –

[3] For example, if the agent discovered that she had merely *misunderstood* some feature of the actual situation, she might regret her action but still regard it as truly her own.
[4] For practical purposes these phrases are equivalent, as I explained in section 2.1.4.
[5] I'll often refer to this case as 'the Good Samaritan case'.
[6] Darley and Batson 1973:107-108. Each of these reactions by subjects are common in situationist experiments. I'll discuss some of the possible explanations and implications in later sections.

was significantly correlated with intervention behaviour (stopping to offer help, or walking past).[7]

If Samantha learned that (a) was a cause of her action, she would not F-accept the action.[8] Therefore (a) is a subverting cause of her action.

3.2.2. Example 2: Social influence and bystander intervention

Martha has volunteered to take part in a market research survey.[9] The researcher hands Martha a questionnaire to fill out, and then goes into the next room where she falls heavily. Despite a loud crash, and the researcher's cries of "Oh my God … my foot … I … I … can't move it. Oh … my ankle", Martha continues to fill out her questionnaire. A cause of this action is:

(b) Another person has volunteered for the survey at the same time, and is sitting next to Martha completing a questionnaire.

If this second volunteer had not been present, Martha would have offered to help the researcher.

This example matches an experiment by Latané and Rodin (1969). When subjects were alone filling out the questionnaire, almost 70% offered to help the researcher. When subjects completed the questionnaire alongside a stooge instructed to act passively, only 7% offered help.[10]

If Martha learned that (b) was a cause of her action, she would not F-accept the action.

3.2.3. Example 3: Influences on social judgements

Judy is assessing the qualities of a candidate for a job vacancy. Judy makes a judgement that the candidate possesses a high degree of flexibility in solving problems. A cause of her action is:

(c) Judy has been told that she will meet the candidate later.

If event (c) had not occurred, Judy would have judged that the candidate had only an average degree of flexibility in problem-solving.

This case matches a study by Nisbett and Bellows (1977).[11] Subjects were asked to read a file about the candidate and to make four judgements: to what degree the candidate was likeable, intelligent, sympathetic to others' feelings, and flexible in problem-solving. Five characteristics about the candidate were varied among different groups of subjects. For example half of the subjects were told that the candidate was physically attractive, while the other half were told nothing about the candidate's appearance. The other varied characteristics were: whether the candidate had an excellent academic record; whether she had spilled a cup of coffee during her interview; whether she had been in a serious car accident which still caused her pain; and whether the subject would later meet the candidate.

---

[7] Darley and Batson 1973:105.

[8] An agent F-accepts her action if she accepts it as her own, and is satisfied with that acceptance. Being satisfied is a matter of having no interest in making changes. See sections 2.1.3 and 2.1.4.

[9] I'll often refer to this case as 'the market research case'.

[10] For a discussion of these and other similar findings, see for example Doris 2002:32-33, Ross and Nisbett 1991:41-44, or Latané and Darley 1970. In section 3.8.5 I'll discuss the experimenters' attempts to explain the subjects' actions in studies like this one and the 'Good Samaritan' case above.

[11] See Nahmias 2001:210-214 for more discussion of this experiment. This study also investigated the accuracy of the subjects' explanations of their own assessments – I'll discuss those findings in section 3.8.1.

Subjects who were told that they would meet the candidate judged her flexibility in problem-solving to be significantly higher than those who did not.[12] This is an example of a well-established tendency:

> people tend to give more favorable ratings on a number of dimensions to people whom they believe they are about to meet than to people whom they do not expect to meet.[13]

If Judy learned that (c) was a cause of her action, she would not F-accept the action. Therefore (c) is a *subverting cause* of her action.

3.2.4. Example 4: Contextual priming

Ingrid enters a room and interrupts a conversation which was already in progress between two other people. A cause of this action is:

(d) A few minutes earlier, Ingrid was completing a word puzzle. Some of the words in the puzzle were connected with a theme of rudeness.

If she had been completing a different word puzzle, Ingrid would not have interrupted the conversation.

An experiment by Bargh, Chen and Burrows (1996) demonstrates this "priming" effect. Subjects were asked to complete a short test in which they had form sentences from sets of scrambled words. After completing that test, they went into a different room to find the researcher, who was engaged in a conversation with a confederate. Of subjects whose puzzle included words connected with rudeness – such as *brazen* and *obnoxious* – more than 60% interrupted the conversation within ten minutes. Fewer than 20% of those whose puzzle contained words connected with politeness made an interruption.[14]

If Ingrid learned that (d) was a cause of her action, she would not F-accept the action.

3.2.5. Example 5: Self-regulation and logical reasoning

Gina gives up trying to solve a difficult problem she has been working on. A cause of her action is:

(e) A few minutes earlier Gina resisted the temptation to eat the chocolate chip cookies which were on the table in front of her.

If Gina had eaten the cookies, she would not have given up working on the problem.

This example matches an experiment run by Baumeister, Bratslavsky, Muraven and Tice (1998). Subjects required to refrain from eating chocolate cookies gave up much more quickly on a complex puzzle task than subjects who ate cookies, or subjects in a control group.

Further related experiments appear to demonstrate that logical reasoning and "self-regulation" (which includes resisting temptations) deplete the same inner resource.[15] For example if Gina had completed the complex puzzle first, she might now fail to resist the

---

[12] Judgements about the candidate's intelligence were strongly correlated with the "academic record" characteristic, as one would expect. However, the subjects' judgements about whether the candidate was likeable were most influenced by whether she had spilled her coffee during the interview.

[13] Nisbett and Wilson 1977:222.

[14] A third group of subjects completed a puzzle with only neutral words, none of which related to rudeness or politeness. Fewer than 40% of these subjects interrupted the conversation. For a discussion of similar effects, see for example Bargh 2008.

[15] These effects are discussed in Baumeister 2008:69-74.

cookies. Alternatively, having earlier resisted the cookies successfully, she might now fail to resist eating a doughnut.

Cause (e) is a subverting cause: if Gina learned that (e) was a cause of her action, she would not F-accept the action.

3.3. IMPORTANT FEATURES OF SUBVERTING CAUSES

I've introduced five examples of actions with subverting causes. It will be useful discuss some of their most important features, since I will draw on these key cases extensively in my arguments. To recap, an action has subverting causes when it is true that:

> if the agent knew all the causes of her action, she would not F-accept the action.

I stipulated that the five agents I described would not F-accept their effective desires, if they learned of the causes of their actions. I think this would be the natural reaction of many people in the same situation, but I do *not* claim that this must be true of everyone in the same situation. Some people performing the same action in the same situation would F-accept their effective desires if they learned of the causes; for these people the causes are not subverting.[16]

3.3.1. Causes: necessary, proximate and relevant

The causes of an action include any proximate, relevant event without which the action would not have occurred. The notion of causal proximity I'm employing here is the one which leads us to exclude events such as Columbus's discovery of a new continent in 1492, or events on the other side of the galaxy, from the set of causes of an action in Britain in 2010.

The notion of causal relevance which I'm relying on is the one which leads us to discount as causes those events which are strictly necessary for an effect to occur, but are merely "in the background". Such events are simply not relevant to the issue under investigation. For example, when the effect is an action and the issue is moral responsibility for it, the beating of the agent's heart and the arrival of photons on her retina are not relevant events.

I won't try to define these notions of proximity and relevance more precisely; but I take it that we already employ these same notions in the everyday judgments we make about causes and effects when attributing moral responsibility.

A cause can include an event *within* the agent's body, such as a brain haemorrhage or the growth of a brain tumour. Such events, if they are causes of actions, can be proximate, relevant and not in the background.

To avoid cumbersome phrasing I won't always include the phrase "proximate and relevant" when I speak of an action's causes, but I will always mean to pick out only those causes which are proximate and relevant.

3.3.2. Knowledge about the cause

The definition of subverting causes invokes a counterfactual scenario in which the agent "knew all the causes of her action". In this counterfactual the agent knows, of each cause, *that the event occurred* and also *that the event was a cause* of her action. I'll use the phrase "to know all the causes of an action" to imply both of these kinds of knowledge about the cause.

However, in the actual situation, the agent may or may not have either kind of knowledge. In all five of my key cases, the agent knows of *the occurrence of* the subverting

[16] I'll discuss an interesting example of this kind in section 4.1.

cause. But a cause can still be subverting if the agent does not know of its occurrence. For example, a subliminal stimulus might be a subverting cause of an action.[17]

In none of my key cases does the agent know that the cause *was a cause* of her action. But it is not essential to a cause's being subverting that the agent does not know it was a cause. For example, suppose an agent knows that she was experiencing a phobia when she acted, and that this was a cause of her action. Here, the known cause is a subverting cause if she *does not* identify with the action – i.e. if she does not F-accept the action as her own. In the counterfactual scenario, her knowledge of the cause is the same as it is in the actual situation. She still would not F-accept the action in that counterfactual scenario, and so the cause is a subverting cause.

3.3.3. Actions, reasons and causes

There are two important and controversial issues about motivating reasons, on which I want to remain neutral.[18] The first is the question 'Are reasons causes of actions?'. The second is 'What is the metaphysical nature of reasons?'. Some philosophers hold that reasons are mental events or states, and that they are causes of actions. Others hold that reasons are not mental but external events or states, and that these are causes of actions. Still others hold that reasons are not causes of actions at all. I think my arguments and the definition of subverting causes can be 'theory-neutral', in the sense that they can be accepted no matter what one's position on these two issues about reasons.

The claim about causes which I need for my arguments is that events which are not mental events or states can be causes of actions. For example, I claim that snowfall overnight can be a cause of Bob's putting on his boots this morning. This statement is compatible with the view that snowfall overnight is a reason for Bob's action. It is also compatible with the view that a reason for Bob's action is that he wants to stay warm and believes that putting on his boots will enable him to stay warm. And, furthermore, it is compatible with the view that Bob's desire to stay warm and his belief that the action will satisfy it are causes (or, jointly, *a* cause) of his action.[19] All that I require, for my claim about causes, is that the snowfall is a proximate and relevant event without which Bob would not have performed the action. In what follows I will usually avoid speaking of reasons for actions, focusing instead on causes of actions.

An action which has a subverting cause may nevertheless be an *intentional* action, either in the sense that it is an action performed *for a reason*, or in the sense that the action has a reason as a cause.[20] For example, a cause of Ingrid's interrupting a conversation is that she completed a word puzzle shortly beforehand – a subverting cause.[21] This does not imply that the action is non-intentional; the action may well be performed for a reason, or have a reason as a cause. Furthermore, the reason may be a normatively good reason – either prudentially or morally good, or both. The presence of a subverting cause does not imply that the action is irrational. On one view of reasons and actions, the reason may be Ingrid's desire to go to meet someone, combined with a belief about how to satisfy that desire. On another view, the reason may be the fact that Ingrid has promised to meet someone. On either view,

[17] A subliminal stimulus is one which cannot reach *conscious* awareness in the perceiver (in this definition I follow Dijksterhuis, Aarts, and Smith 2005). Nevertheless, subliminal stimuli can be perceived unconsciously. In a study by Karremans, Stroebe and Claus (2006), presenting subliminal images of the brand name of a drink influenced subjects to choose that brand if they were already thirsty.

[18] By "motivating reasons" I mean, broadly, reasons which actually motivate agents to act. These are often distinguished from "normative reasons" which are (broadly) reasons in favour of acting, or which justify acting.

[19] On a Davidsonian account, the desire and belief pair make up *the primary reason*, which is a cause of the action (Davidson 1963).

[20] The latter is Davidson's definition of an intentional action.

[21] See section 3.2.4.

the presence of the reason is compatible with the presence of a subverting cause. Ingrid's moral responsibility is subverted if she would not F-accept the action as truly her own, if she knew all of its causes.

3.4. MY "MAIN ARGUMENT" AGAINST FRANKFURT

My argument against Frankfurt involves three of the statements which I used to define his position in section 2.2, namely:

(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).[22]
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility (A→R).

I will argue that (5) is false. From that, it follows that (6) is false if (4) – which is an identificationist axiom – is true.[23] If F-acceptance is not sufficient for an action being truly the agent's own, then F-acceptance is not sufficient for DMR for the action either (given that being truly the agent's own is necessary for DMR).

My argument to show that (5) is false proceeds as follows. Using the five key cases discussed above, I aim to demonstrate that the agent's F-acceptance of an action (Frankfurt's identification condition) is not sufficient for the action's being truly her own. In each of the five cases, I claim that the agent does F-accept her action, but that the action is not truly her own because the action has a subverting cause.

I'll defend each part of my argument separately. In the next section (3.4.1), I'll argue that the agent does F-accept her action, in each of the five key cases. Then in section 3.4.2 I'll argue that the action is not truly the agent's own, in each of the five cases. From this it follows, given (4), that the agents are not directly morally responsible for these actions. In section 3.4.3 I'll consider some objections to the conclusion that the agents in the five cases are not DMR for their actions.

It's important to note that the conclusion is true if the premises are true for *any* action. To refute the argument, an objector would need to show, for *all five* key cases, that one of the premises is false. A single counterexample will be enough to establish my argument.

I'll also defend my claim that the events (a) to (e), which I picked out in my descriptions of the five key cases, can indeed be causes of actions. Support for this claim in each case comes from the study which matches the action. However, five examples may not be enough to convince a doubter that such phenomena are genuine. Therefore, in section 3.6, I'll discuss a much wider range of experimental evidence for similarly subverting causes of actions. In sections 3.7 and 3.8 I'll defend my conclusions from that evidence against objections. This should establish my claims about the causes of the five cases as convincing.

Furthermore, even if there are never in fact any subverting causes of actions, my argument nevertheless reveals something important about the conditions for direct moral responsibility. Simply considering the potential threat posed by subverting causes reveals that Frankfurt's identification condition (F-acceptance) is not sufficient for the action being truly the agent's own, and so not sufficient for direct moral responsibility, even if subverting causes do not in fact occur. The F-acceptance condition might in practice be perfectly correlated with

[22] An agent F-accepts her action if she accepts it as her own, and is satisfied with that acceptance. Being satisfied is a matter of having no interest in making changes.
[23] If ¬(A→O) while (R→O), then it follows that ¬(A→R).

DMR (setting aside conditions about history and normative competence), but it would not be logically sufficient for direct moral responsibility.

But I aim to go further, and show that there are actual cases in which agents F-accept their actions but are not DMR for them, because the actions are not truly the agents' own.

3.4.1. Subverting causes and F-acceptance

I'll begin by defending this claim: in each of the five key cases, the agent F-accepts her action, and hence Frankfurt's identification condition is met. In effect I have stipulated this as a fact about the agents in the five cases. But for my argument to be convincing it must also be plausible that, in five real-life cases matching the ones I've described, the agent F-accepts her effective desire to act, and so F-accepts her action.[24] This *is* plausible, because the cases are representative of actions that all of us perform every day. Agents who perform actions like these usually accept them as their own, and are satisfied with that acceptance, in the Frankfurtian sense of those phrases. None of the actions would seem remarkable, were it not for our knowing about the influence of the subverting cause.

It may seem that, in some of the cases at least, the agent's *lack of reflection* upon her own effective desire might prevent her F-acceptance of it. But on Frankfurt's final account, F-acceptance requires only that the agent be *satisfied* with the effective desire, in the sense that she has "no interest in making changes". It's not necessary that the agent knows in advance of the action that she has that desire.[25] Nor is it necessary that the agent reflects on the desire after the action. For example, if Phoebe crosses the road to avoid some workmen on the pavement, all the while talking on her mobile phone, she may not at any time reflect on her effective desire to cross the road. Nevertheless, according to Frankfurt, she may F-accept that desire and be DMR for the action. Thus, F-acceptance is not precluded either by lack of awareness of, or by lack of reflection upon, one's effective desire.

It might be claimed that the agents, in some of the examples, must feel uneasy or conflicted in some way about their actions – and that this prevents their satisfaction and F-acceptance. The thrust of the objection would be that feeling uneasy or conflicted about one's action indicates a lack of satisfaction with the effective desire. In reply I make two points.

The first is that some of the agents in the examples might not feel at all uneasy. For example I think it's very plausible that Judy does not feel uneasy about her judgement of the job candidate's flexibility; the consequences of her action are not very grave. In the situations described in the other examples, it's easier to imagine agents who do feel conflicted about their actions. But, at the same time, I think it's still very plausible that *some* agents in those situations do not feel conflicted at all. Even if they reflected further on their actions, *without of course learning about the subverting causes*, they still would not feel conflicted. In section 3.8.2 below I'll discuss the phenomena of dissonance and confabulation, which may help to explain why some agents do not feel conflicted, even after an action with a subverting cause. But even without dissonance and confabulation effects, some agents may not feel conflicted if they simply do not reflect very carefully on their actions. For my argument, I need only a single instance of an agent who F-accepts an action with subverting causes.

The second point I can make in reply to this objection is that feeling uneasy or conflicted about an action does not prevent identification with the effective desire.[26] We face many difficult decisions in everyday life, and we often feel conflicted about what to do, and

[24] F-acceptance of the effective desire is equivalent to F-acceptance of the action – see section 2.1.4.

[25] In fact Frankfurt holds that it is possible to be satisfied with, and so identified with, an unconscious desire (1992:99 and 1971:13).

[26] I have in mind Frankfurt's theory here of course, but this is likely to be true on any plausible account of identification.

uneasy about whether we made the right choice. But that does not entail that we do not identify with the course of action we take, even if it follows a lot of anxious deliberation. After all, identification with one desire rather than another is one of the hallmarks of a successful conflict resolution. It is not always followed by a thoroughgoing inner peace.

3.4.2. Subverting causes and being truly the agent's own action

I'll now defend my claim that, in each of the five examples, the action is not truly the agent's own. I make two points here.

The first point is about knowledge. There are two scenarios in which the agent may or may not F-accept her action. The first of these is the actual situation, in which she does F-accept her action – as I discussed in the previous section. The second is the counterfactual scenario in which the agent learns of the subverting cause of the action, and does not F-accept it. The key question is: in which of these scenarios does her F-acceptance, or lack of it, best indicate whether the action is truly her own?

The answer is: the counterfactual scenario. All of the events preceding the action are common to both the counterfactual scenario and the actual situation. But in the counterfactual scenario the agent has more knowledge, and that knowledge is relevant to her F-acceptance. If asked, the agent would prefer to have more knowledge (ceteris paribus). With more knowledge, her F-acceptance reflects more reliably whether the action is truly her own. If she does not F-accept it in that counterfactual scenario, then the action is not truly her own. I think followers of Frankfurt should accept this argument, since it employs the central elements from his own theory in a way which is consistent with the theory's objectives.

My second point is that, in every example, knowing the role of the subverting cause, I would not judge that the action is truly the agent's own. Nor would most of the people with whom I've discussed the issue. Nor would I feel that the action was truly *my* own, if I were an agent in the same circumstances *and I learned that the subverting cause was a cause of my action*.

In each case, without the presence of the subverting cause, the agent would have performed a different action.[27] In each case, had she learned that the subverting cause was a cause of her action, she would not have F-accepted the effective desire as her own. This shows that the effective desire, and so too the action, are not truly her own.

I stipulated as a fact about the agents in the five key cases that they would not F-accept their actions, if they knew all the causes. I did not claim this would be true of *everyone* in the same situation. A single example in which it's true is enough to support my argument against Frankfurt.

3.4.3. Subverting causes and direct moral responsibility

I have argued that statement (5) of Frankfurt's position is false: F-acceptance of an action is *not* sufficient for its being truly the agent's own. I combine this finding with the identificationist axiom (4) that, for direct moral responsibility for an action, the action must be truly the agent's own. From these two statements it follows that, contra Frankfurt, F-acceptance is *not* sufficient for direct moral responsibility.[28]

An obvious objection is that it seems wrong to excuse the agents in these five cases from DMR. It may well be, of course, that all of the agents did in fact feel themselves directly morally responsible for their action, since *they* did not know that the unendorsed cause was a

[27] This follows from the definition of a cause in section 3.1: any proximate, relevant event *without which the action would not have occurred*.

[28] If $\neg(A \rightarrow O)$ while $(R \rightarrow O)$, then it follows that $\neg(A \rightarrow R)$. Therefore statement (6) is false.

cause. But that is not enough for DMR, on the identificationist view that the action must be truly the agent's own.

It's also true that we could hold the agents *indirectly* morally responsible for their action – via the tracing principle. For example, suppose we hold Gina indirectly morally responsible (TMR) for giving up working on the problem after resisting the cookies.[29] We might think that some earlier action, for which she is DMR, led to her giving up. Perhaps yesterday she noticed that she got distracted easily from her work, after sitting at a table where there were chocolate cookies in front of her. She resolved to put the cookies out of sight in future. But today, she chose again to work at the same table with the cookies in plain view, and resisted them once more. It might then be reasonable to hold Gina TMR for giving up today.[30] But even if we can hold Gina TMR for her action, this would not provide an objection to my conclusion, which is about *direct* moral responsibility.

Moreover, the fact that Gina *might* be TMR for her action, because of some earlier action, helps to explain some of our intuitive *reluctance to excuse her* from moral responsibility. When we feel intuitively that someone is morally responsible for an action, we don't usually distinguish DMR from TMR. The feeling generally comes before we begin the process of working out why the agent is morally responsible – if indeed that process occurs at all. When an everyday action has a subverting cause, we may instinctively feel that the agent is morally responsible. But I suggest that feeling arises because we are so confident of our judgements about moral responsibility for similar everyday actions, in which we ignore subverting causes. After thinking carefully about whether actions with subverting causes are truly the agents' own, I believe the most common reaction will match mine. More importantly, I think followers of Frankfurt should accept my conclusion, for the reasons I gave above. Once again, a single counterexample is enough to establish my argument.

Some other objections spring easily to mind. It's possible that – in some of the examples at least – the agent would feel *herself* to be directly morally responsible, *even if she were told* of the role of the subverting cause. A further objection employs the claim that the agent, in each of the five cases, is DMR because she *is blameworthy* for performing the action. These objections are not open to followers of Frankfurt, if they accept the identificationist axiom (4).[31] Nevertheless I will respond to these objections. I'll delay doing so until chapter 4, because my replies will draw on experimental evidence about subverting causes which I have not yet discussed.

3.5. MY POSITIVE PROPOSAL: THE CFA CONDITION

In the last section I argued, against Frankfurt, that F-acceptance of the effective desire is not sufficient for direct moral responsibility for the agent's action.[32]

I now offer a positive proposal.[33] I suggest that the following condition is necessary for direct moral responsibility:

> *CFA* The agent would F-accept the effective desire, at the time of acting, if he knew, at the time of acting, all the proximate and relevant causes of

[29] See section 3.2.5.

[30] Note that, on the tracing conditions discussed in section 1.2, Gina would bear traced moral responsibility only if the resultant occurrence of her later uncontrolled action was reasonably foreseeable by her at the time of an earlier action for which she was DMR. I'll say more about whether this is plausible in section 7.1.2.

[31] Axiom (4): for direct moral responsibility for an action, the effective desire must be truly the agent's own.

[32] An agent F-accepts his effective desire if he accepts it as his own, and is satisfied with that acceptance, where being satisfied is a matter of having no interest in making changes. See section 2.1.3.

[33] My CFA condition is partly inspired by Nozick's notion of "acts in equilibrium" (1981:348-352), though Nozick himself does not use it as part of a condition of freedom or moral responsibility.

> that desire's being effective (except in so far as any additional knowledge gained gives him new information about means to satisfy his desires).

After making three simplifications, a working version of this definition is:

> The agent *would F-accept his action, if he knew all the proximate and relevant causes* of the action.

The first simplification is to substitute the action for the effective desire as the object of F-acceptance; these are equivalent for practical purposes.[34] The second simplification is to remove the clause in parentheses. This clause is designed to exclude knowledge which would make the agent *regret* his action, yet still feel that the action is truly his own; I'll discuss it in section 4.2. The third simplification merely removes the two clauses which specify that the agent's counterfactual knowing and F-acceptance would occur at the time of acting. In what follows I'll usually avoid the precise but clumsy phrasing in favour of the simpler definition.

The definition of CFA is based on Frankfurt's identification condition, since his is the identificationist account I am disputing in my main argument. The important change I'm adding is that the F-acceptance required of the agent is now counterfactual. It seems likely to me that any identificationist theory of moral responsibility will need to incorporate a similar counterfactual condition, though I won't argue for that claim here.

The phrase "if he knew all the proximate and relevant causes" stands in need of clarification. The knowledge the agent has in the counterfactual scenario is more than simply knowledge of the *existence of* the causes of his action. He knows also *that they are causes* of his action. However, he need not understand *how it is that* these events are causes of his action: he need not understand all the causal connections and intermediary processes, nor be able to predict other scenarios in which the causes would be effective. Henceforth I'll use the phrase "to know the causes" in this way.

To recap, I'm focusing on three statements defining Frankfurt's position:

(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility (A→R).

In section 3.4 I argued that (5) is false, from which it follows that (6) is false if (4) is true. I now propose that the following statement is true:

(7) The effective desire is truly the agent's own only if the CFA condition is met (O→CFA).

From (7) and (4) it follows that:[35]

(8) The agent is directly morally responsible for an action only if the CFA condition is met (R→CFA).

Now, the following is also true:

(9) It is not the case that F-acceptance of an effective desire is sufficient for the CFA condition to be met (¬(A→CFA)).[36]

[34] See section 2.1.4.
[35] If (O→CFA), while (R→O), then (R→CFA).

From (9) and (8) we can also derive that (6) is false, which was my main conclusion in section 3.4.[37]

I don't claim that the CFA condition is *sufficient* for DMR, or for the effective desire's being truly one's own. Nor do I claim that the CFA condition, in conjunction with Frankfurt's identification condition, are *jointly sufficient* for DMR (or being truly one's own). In fact, I think that conditions concerning the agent's history and normative competence are also necessary for an action's being one's own, and for DMR.[38] I argue only that CFA is necessary for direct moral responsibility, and that followers of Frankfurt should agree with me about this.

3.6. THE EXPERIMENTAL EVIDENCE

I've already mentioned some experimental evidence that the CFA condition is sometimes not met, in the discussion of the five examples of subverting causes in section 3.2. This limited evidence will not be convincing without an appreciation of the wider pattern of research into which these experiments fit. In the remainder of this chapter I'll give an overview of that research. I have two specific objectives in doing so. First, I want to support the key premise of my main argument in section 3.4, by establishing that:

(A) There are actions with subverting causes.

Second, I want to establish a stronger claim that I'll use in my later "sceptical argument" in chapter 7:

(B) There are many everyday actions with subverting causes.

I don't aim to mount a conclusive defence of either claim. I aim only to show that each one is extremely plausible. To support claim (A), I will demonstrate that:

(A1) The five examples of subverting causes in section 3.2 are not isolated incidents which occur because of dubious experimental methods. Rather, they are representative of a large body of credible experimental evidence in mainstream fields of psychology.

To establish that (B) is plausible, I will defend these subsidiary claims:

(B1) There are good grounds to believe, on the basis of existing experimental evidence, that many everyday actions have subverting causes.

(B2) There are good grounds to believe that there are many more subverting causes of everyday actions as yet undiscovered by experimenters.

Claims (B1) and (B2) are interrelated. Both are matters of degree. The more plausible is claim (B1), the less support is required from claim (B2), in order to establish (B). The more plausible is claim (B2), the less support is required from claim (B1).

Over the next several sections I'll present evidence which establishes claims (A1) and (B1) as extremely plausible. Finally, in section 3.8.6, I'll review that evidence and give a brief inductive argument in favour of claim (B2).

---

[36] Some agents who do F-accept their effective desires, would not F-accept them if they knew all the causes – I argued for this in section 3.4.1.

[37] From (R→CFA) and ¬(A→CFA) it follows that ¬(A→R), i.e. that (6) is false. From (7) and (9) we can also derive that Frankfurt's (5) is false: if (O→CFA) and ¬(A→CFA), it follows that ¬(A→O).

[38] See sections 2.3.1 and 2.3.2.

I'll begin by discussing 'situationist' psychology, the field in which three of the effects in my five key cases were discovered.[39]

3.6.1. Situationist social psychology

Researchers in the 'situationist' field of social psychology have, over the last sixty years, been amassing evidence that we make a "fundamental attribution error" when explaining human actions. We often operate with

> [an] inflated belief in the importance of personality and character traits, together with [a] failure to recognize the importance of situational factors in affecting behavior.[40]

The research offers support to the idea that we do, in practice, weigh an agent's personality and character traits very heavily when explaining and predicting her behaviour. At the same time, though, the research reveals that such explanations and predictions are often inaccurate. The examples I discuss are only a tiny sample of that research.[41]

Philosophers have perhaps been a little slow to address the implications of situationist research.[42] Most of the (limited) philosophical engagement with situationism has focused primarily on the implications for virtue ethics, and the influence of the agent's character on his ethical behaviour. Few philosophers have addressed the implications of situationist research for free will and moral responsibility. The most notable exceptions are Nahmias, Doris and Nelkin.[43] Nahmias and Doris propose positive theories which take account of the findings from situationist research. I address their theories in sections 5.1 and 5.2.

It's important to stress at the outset of my discussion that I am not claiming that character traits do not exist. Nor am I claiming that character traits never help to explain or predict agents' actions. Few (if any) situationist psychologists would make such strong claims. Nor do I need to claim that personality or character traits are only rarely causes of actions; nor that they are causes less often than situational factors.[44] In fact my argument has little to do with personality and character traits at all.

I'll briefly describe two important situationist studies, in order to give an impression of the scope of the effects which have been discovered.

One of the most striking situationist experiments was first performed by Stanley Milgram in 1960. The experiment has been repeated many times, by Milgram and by other researchers, sometimes with small variations from the original. A representative description is

---

[39] To recap these examples very briefly: Samantha walks past a slumped man after being told that she is running late for her presentation; Martha ignores the cries of a researcher because someone is sitting passively next to her; and Judy judges that a job candidate possesses a high degree of flexibility after being told that she will meet the candidate later. See sections 3.2.1 to 3.2.3.

[40] Ross and Nisbett 1991:4.

[41] My discussion draws heavily on Ross and Nisbett's classic 1991 'textbook' of situationism, and also on the lengthy reviews of situationist research in Nahmias 2001 and Doris 2002. My aim here is merely to reflect the key themes and issues, which I will cover in much less depth than these authors.

[42] Doris comments that "moral philosophers have not typically engaged [with] empirical psychology in anything remotely approaching the depth required to see exactly where and how it might matter for ethics" (2002:4). Even Milgram's extraordinary findings (discussed below), which have obvious ethical significance, have been little discussed by philosophers – see Doris 2002:39.

[43] Nahmias (2001) discusses the problems for *free will* posed by situationist research. Doris (2002 ch.7) discusses the problems posed for responsibility. Nelkin (2005) discusses both free will and responsibility, but does not propose a positive theory of either.

[44] Whether these claims are true is likely to depend on the breadth of the notion of personality and character traits being employed. Situationism "denies that people typically have *highly general* personality traits that effect behavior manifesting a high degree of *cross-situational consistency*" (Doris 2002:38-39, my italics).

as follows.[45] A stooge confederate of the research team plays the role of a volunteer who is trying to learn a series of word associations (the "learner"). Under the guidance of an "experimenter" (often wearing a lab coat), the subject is instructed to administer an electric shock to the learner every time the latter gives an incorrect answer to a question.[46] The intensity of the shocks increases as the task continues, so that eventually the subject dispenses shocks labelled on the equipment as "375 volts" and "Danger: severe shock". The stooge learner does not actually receive the shocks, but nevertheless acts convincingly by appealing more and more urgently for the shocks to stop, pleading that he should be excused because of a heart condition, and finally falling silent as the voltage reaches the most dangerous levels. A typical result of the experiment is that about two-thirds of subjects will continue to administer shocks all the way to the very highest levels.[47]

The conclusion drawn by situationists is that the subjects are influenced to a remarkable degree by certain factors in the situation, rather than that they are displaying a remarkable degree of callousness in their characters. Some of these factors, when varied, correlate strongly with the number of compliant subjects. For example in one study, when the "experimenter" was in a different room and gave instructions by telephone, only 21% of subjects administered the shocks to the end.[48]

Another totem of situationist literature is the 'Stanford Prison Experiment' run by Philip Zimbardo in 1971.[49] Twenty-one volunteers were divided randomly into two groups at the start of the experiment – "prisoners" and "guards". The intention was to run a simulation of a prison for two weeks, using rooms in an empty part of Stanford University's psychology department as the "cells". The simulation was terminated after six days, following what can be fairly described as "a precipitous descent into barbarism".[50] The guards behaved sadistically. The prisoners rebelled, and one staged a hunger strike. Five prisoners had to be released early from the experiment because they were acutely distressed. Even a former convict playing only a minor role in the simulation, as a "parole board chairman", found himself sickened by his own actions.[51] These extreme effects are all the more startling when we consider that all the participants knew they were taking part in a time-limited simulation, from which they could withdraw at any time, and which was not especially realistic in many ways – there was no physical punishment of prisoners, the "cells" were ordinary rooms, and there were no high walls around the university.

It is difficult to draw precise conclusions from the Stanford Prison Experiment, because – unlike the other experiments I've described – there were many situational factors at work simultaneously, and no control conditions against which their influence could be compared. Nevertheless it seems reasonable to conclude that, when many situational factors work together, the effects can be very significant.

3.6.2. Influences processed automatically

Meanwhile psychological research is revealing that our actions can be influenced in very surprising ways by influences which are processed automatically or unconsciously. Many disparate effects have been discovered, which are not easily summarised by a single

[45] For much more detailed discussions of Milgram's experiment, see Ross and Nisbett 1991:52-58 or Doris 2002:39-51.
[46] No coercive pressure is applied; although subjects are given insistent instructions by the "experimenter" in the lab coat, they are free to desist at any point without threat of sanctions.
[47] Doris 2002:45.
[48] Doris 2002:46.
[49] Discussed in Doris 2002:51-53. For the fullest (astonishing and unforgettable) account of the experiment, see Zimbardo 2007 chapters 2 to 11.
[50] Doris 2002:53.
[51] Zimbardo 2009.

description or label. Moreover, it is too early to expect comprehensive theories to explain these effects, since many of them have only recently been researched (in the last decade).

In section 3.2.4 I discussed one example of *contextual priming*, in which Ingrid interrupts the experimenter after completing the rudeness-themed word puzzle. This particular effect seems to work because Ingrid's *mood* is influenced by the content of the priming event (the words in the puzzle). Other studies have shown that agents' *goals* can be primed too, sometimes even subliminally (so that the subjects are not consciously aware of the existence of the priming event). For example, a 2001 study found that

> subliminal priming of the goal of co-operation causes participants playing the role of a fishing company to voluntarily put more fish back into a lake in order to replenish the fish population, compared to a control condition.[52]

It may be that these effects confer a benefit to us as individuals by helping to ensure that our moods, goals and actions are well matched with those of our social group, which we pick up through a host of subtle cues which need not be processed consciously. Such explanations are somewhat speculative, though, since this area of research is still relatively young.

In one of my key cases, Gina gives up on a problem after resisting an opportunity to eat chocolate cookies (an act of self-control or *self-regulation*).[53] A series of similar experiments reveal that

> logical reasoning in the service of making deliberate, conscious choices appears to deplete the same resource used for self-regulation.[54]

A speculative explanation for this state of affairs is that a powerful but energy-intensive neurological system evolved to enable *one* of these functions in humans (perhaps it was self-regulation – the ability to override automatic behavioural impulses). This same system was later adapted to perform a *second* function (perhaps logical reasoning and decision-making evolved later than self-regulation).

A more famous experiment shows how *automatic arousal* processes can influence our actions in unexpected ways.[55] The subjects were single men in a park who had been given a spurious explanation of the purpose of the study. The attractive female researcher offered to explain the study in more detail when she had more time, and gave them her phone number. One group of subjects met the woman on a footbridge over a deep gorge which swayed in the wind; a control group met her beside a park bench. Of the men on the bridge, 57% called to speak to the researcher again; only 30% of the control group did so. The apparent explanation is that the men on the swaying bridge were physiologically aroused by fear or exhilaration, and found the researcher more attractive as a result:

> By failing to recognize why they were aroused, people were more attracted to someone than they would otherwise have been.[56]

Needless to say, the subjects did not suspect that the location of the meeting might have influenced their action.

[52] Bargh 2008:143.
[53] See section 3.2.5.
[54] Baumeister 2008:73.
[55] Dutton and Aron (1974). For a discussion of this study, see for example Wilson 2002:100-102.
[56] Wilson 2002:102.

Meanwhile studies of *automatic attitudes* show that our preferences and choices can be influenced by some very surprising factors. It seems that automatic processes provide evaluations of stimuli in our environment, and that these evaluations can influence our actions in ways which we do not recognise. For example a 2005 study found significant correlations between voter ratings of the competence of political candidates and the facial appearance of those candidates.[57] Other recent studies have found that

> compared to what you'd expect by chance alone, there are more people named Ken who moved to live in Kentucky, Florences who moved to Florida, and more named Louis who moved to St. Louis; there are more Dennises and Denises who become dentists and Lauras and Lawrences who become lawyers, compared to people with names that do not share letters with these occupations.[58]

In a 2006 study, student subjects read a fictitious newspaper article about another student who had won a prestigious award in mathematics. Some subjects read that they shared their birthday with the fictitious award winner. At the end of the semester, those subjects gained higher average grades in mathematics than students in the control group – presumably because their studying activities were influenced by reading the article.

These different kinds of study have in common the finding that actions can be influenced by unexpected factors because of the mediating role of automatic processes. Such an influence is a cause of an action if the agent would not have performed the action in the absence of that influence. It's clear that the influences in these studies are causes in that sense, at least for many of the subjects in each experiment. It seems to me very likely that many of the agents in these studies would regard the causes of their actions as subverting: if they knew of these causes, they would not F-accept their actions.

3.7. OBJECTIONS TO MY CLAIMS ABOUT SUBVERTING CAUSES

I'll now discuss some objections to the claims I introduced at the start of section 3.6, which were:

(A) There are actions with subverting causes.

(B) There are many everyday actions with subverting causes.

Those claims are supported by the experimental evidence which I've reviewed in the previous two sections. My discussion of the objections and responses will be similarly brief.[59] My aim is to establish that the claims above are very plausible, not to provide a conclusive defence of them.

3.7.1. The claims contradict common sense

The first and most obvious objection is that these claims, and often indeed even the very *results* of the experiments, contradict common sense. In other words they contradict the way we habitually think about, and what we think we know about, human actions. That it contradicts common sense is not a very convincing objection to an experimental discovery. All of the sciences, including psychology, have revealed truths which contradicted previous common sense. It's particularly difficult, though, for us to accept challenges to our

[57] The studies in this paragraph are discussed in more detail in Bargh 2008:136-140.
[58] Bargh 2008:139.
[59] Many of these issues are covered in greater detail by Doris (2002 chapters 1 to 3) and Nahmias (2001 chapter 3).

commonsense view of *action*. Action is a phenomenon with which we feel intimately acquainted. At least in familiar everyday actions, we believe that we know the causes and that they are not subverting. We need more than experimental results to convince us that such important beliefs are mistaken. What we need, as well, are explanations of how and why we form these mistaken beliefs. As I'll discuss below (section 3.8), psychologists are beginning to provide these explanations.

Meanwhile, it's important to stress that the experiments I have discussed provide only a tiny sample of the research findings in each relevant field of psychology. I selected, for the examples in section 3.2 on which I built my argument, experiments which demonstrate striking effects which can clearly be attributed to a single variable. But any one of the experiments I have quoted could later be discredited without undermining my confidence in the claims I made above, if the research in the broader field remains sound.

In case my claims may seem too strong, it's worth contrasting them with some claims that I'm *not* making. Take the example of Samantha who walks past the slumped man on her way to give a talk on the Good Samaritan.[60] I'm not claiming that *everyone* in the same situation as Samantha would walk past the slumped man after being told of their lateness for a talk. Nor am I claiming that *most* people would walk past the slumped man.[61] Nor am I claiming that being told of their lateness is the variable which correlates *most strongly* with that action.[62] All I need for my claim (A) is that the subverting cause can be a cause of the action, for at least one person. I think it would be implausible to deny that claim. To support claim (B), I need only that there are many actions with similarly subverting causes.

3.7.2. The evidence is gathered in artificial conditions

A second objection is that since the experimental results occur in artificial conditions, my claims are applicable only to actions in such conditions, and not to everyday actions. I think this is implausible. It's true that the experiments which match my examples in section 3.2 were conducted by academic researchers, with subjects who know they have volunteered for some kind of research (though they often do not understand the real purpose of the experiment until after the results have been captured). However, there are many experiments with similar results in which the subjects are not volunteers, and are unaware that their actions are being studied.[63]

Meanwhile in many studies there is also evidence that, even though they know that they are taking part in a research experiment, the subjects treat their decisions and actions seriously. For example, in Milgram's experiments described above, the subjects often displayed signs of great and genuine distress (yet most continued to obey the experimenter).[64]

For less dramatic actions than administering electric shocks, such as giving up on problems, interrupting people in conversations, and expressing judgements about interview candidates, I think it's even more plausible that the same subverting causes can operate in their everyday equivalents. The best reason to think otherwise is simply that we do not *expect* these actions to be influenced by subverting causes. But that very fact also helps to explain why we do not notice their influence. We don't notice that actions like these correlate with factors like the availability of cookies or the content of a word puzzle. We don't notice the correlations *precisely because* these factors seem to be inconsequential.

---

[60] See section 3.2.1.

[61] Of course, in the experiment conducted by Darley and Batson, most subjects in Samantha's situation *did* walk past the slumped man.

[62] In fact, Darley and Batson also tested for correlations with a personality measure (tested by questionnaire), and with the topic of the talk (which was varied for some subjects). The only factor which correlated significantly with the subjects' actions was being told whether they were late. See Darley and Batson 1973:104-106.

[63] A famous situationist example is the 'dime in a phone booth' experiment described in section 4.1 below.

[64] Ross and Nisbett 1991:55.

3.7.3. The results must reflect character differences in subjects

A third objection is that the experimental results must reflect *personality* or *character* differences in the subjects, rather than (or perhaps as well as) the situational factors, priming events and so forth which the experimenters identify as causes.[65] It's important to acknowledge that, in the fields I've discussed, most experiments test a single optional action (e.g. the subject administers the shock, or refuses to do so) against a series of potentially correlated factors (e.g. the subject meets or does not meet the person being shocked; the experimenter wears a white lab coat or casual attire; the experiment is held in a university psychology lab or a dingy basement).

This kind of experiment cannot test whether the subject's action in the experiment correlates with any of her other actions. For example, we cannot say whether subjects who administer the full set of shocks in a Milgram-style experiment also walk past people slumped in doorways more often than subjects who refuse to administer all the shocks. To compensate for this practical limitation, experimenters often test for correlations between subjects' actions and their age, sex, socio-economic background, and reported personality measures – all of which I'll call "dispositional" factors. One of the most important general lessons to be drawn from situationist research is that dispositional factors correlate much less strongly with observed behaviour than the subjects themselves would expect.[66] Sometimes the experimenters find no significant correlations with dispositional factors.[67] Furthermore, there is a very large number of experiments in which dispositional factors correlate much less strongly with subjects' actions than do situational factors.[68] This, of course, is *more* than we need in order to say with confidence that the situational factor is a cause of the action for some of the subjects. *Any* significant correlation between situational factor and observed action would provide us with confidence that the situational factor is sometimes a cause of the action.

We expect – naively, according to situationists – that dispositional factors correlate much more strongly with actions than situational factors. If that were true, we should expect to find agents performing similar actions in widely differing situations. In fact there have been relatively few studies which test consistency of subjects' behaviour in several different situations.[69] One classic study tested a group of school children in a range of situations in which they could choose whether to act honestly.[70] For example they had the opportunity to steal money from an empty classroom, to lie to help a fellow pupil, and to cheat in various different ways in tests. Relying on common sense, one would predict a high degree of cross-situational consistency, reflecting the character trait of honesty. In fact, subjects behaved very consistently across situations which were very similar indeed. They were much less consistent then expected across situations which were only somewhat similar. For example:

> A particular form of cheating, such as copying from an answer key, might correlate strongly (.70) with copying from a key on a similar test at a later

[65] Doris discusses this objection in some detail in relation to the Milgram experiments (2002:47-49).

[66] I'll discuss some of the evidence for this point in section 3.8.1 below. See also Ross and Nisbett 1991 chapters 1 and 5.

[67] See for example Latané and Darley 1970 chapter 12.

[68] See Ross and Nisbett 1991 chapter 1 for an insightful discussion of correlations and "effect sizes".

[69] Such studies are complex, expensive and time-consuming – see Ross and Nisbett 1991:98.

[70] For a more detailed discussion of this 1928 study of honesty by Hartshorne and May, see Doris 2002:24 and 62-64, and Nahmias 2001:197-198. For studies of the consistency of other dispositional traits, see Ross and Nisbett 1991:96-100.

> date, but not correlate strongly (.29) with another form of cheating, such as continuing to work on a speed test after time is called.[71]

Meanwhile experiments which compare actual behaviour with self-reports, or peer assessments of personality traits, typically find much lower correlations than would be expected using common sense.[72]

However, it is important to repeat that I do not need to claim, and indeed I do not claim, that dispositional factors never correlate strongly with actions. Nor do I claim that situational factors usually correlate more strongly with actions than do dispositional factors. My claims (A) and (B) are about the existence and prevalence of subverting causes. In the example cases in section 3.2, and in experimental scenarios like those in the Milgram and Zimbardo studies, what matters to my argument is not the presence or absence of correlations between actions and dispositional factors. What I claim is that the agents of some of these actions would not F-accept their effective desires, if they knew all the causes of the actions. It may well be that there is some correlation between the agent's propensity to perform the action and a dispositional factor. But that relationship is not relevant to the truth of my claims.

I do think it is plausible to conclude, from situationist research, that situational factors *sometimes correlate more strongly* with actions than dispositional factors. But in fact I do not need even that conclusion in order to support my claims (A) and (B) – which are strictly claims about the prevalence of subverting causes.[73]

Hence, while I think the assertion made in this objection is implausible, it does not in fact threaten my argument.

3.8. ERRORS IN OUR EXPLANATIONS OF ACTIONS

I have defended the following claims:

(A) There are actions with subverting causes.

(B) There are many everyday actions with subverting causes.

I've discussed some of the positive evidence which supports (A) and (B), and dealt with some objections to drawing them as conclusions from the evidence.

The claims are difficult to believe, and the positive evidence for them is so surprising, because we think we know why we act. We think we know the reasons for our actions, and we think we know the causes of our actions. There is a conflict between those strongly-held intuitions and the evidence I've discussed above. In such conflicts, we are usually inclined to doubt the evidence more readily than to doubt our intuitions.

To come to believe that the fault lies in our intuitions, we need something more: an *explanation* of how our intuitions often come to be mistaken. In this section I'll discuss several parts of that explanation, together with an overview of the evidence supporting them. Some of that evidence has been discovered relatively recently, and I suspect the explanation will become more and more plausible with further advances in research.

I won't discuss the evidence in great detail: my objective is simply to give an overview of the research in each field, and to show why it helps to make plausible my claims (A) and (B). I'll begin with evidence from situationist experiments that we do make mistakes when we predict and explain actions.

---

[71] Doris 2002:64.

[72] See Ross and Nisbett 1991:98-9, or Doris 2002:63.

[73] By way of comparison, Doris makes much stronger claims than my (A) and (B). For example: "In very many situations it looks as though personality is less than robustly determinative of behavior. To put things crudely, people typically lack character" (2002:2).

3.8.1. Errors in explanation and prediction of actions

In a study by Nisbett and Bellows (1977), college students were asked to assess someone (a fictional candidate called 'Jill') for suitability for a job.[74] The assessments were based on a file of written reports about her application and a prior interview. Jill was rated on four criteria: intelligence, flexibility, sympathy and likeability. Five factors in Jill's file were manipulated: her academic credentials, whether she was described as attractive, whether she spilled her coffee during the interview, whether she had recently been involved in an accident, and whether the subject would later meet Jill.

One group of subjects (the "actors") read the file, gave ratings of Jill, and also reported on which factors had influenced those ratings. A second group ("observers") did not view the file on Jill, but merely estimated which factors would influence them if they were making such ratings.

The "actors" reported that Jill's academic credentials were the factor which most influenced their rating of Jill's intelligence, and indeed there was a strong correlation between those credentials and ratings of intelligence. However, their reports were much less reliable predictors of the other three ratings. For example, "actors" reported that academic credentials also had most influence on their rating of her flexibility; in fact by far the most important factor in that rating was whether they were told that they would later meet Jill. Perhaps the most interesting result was that

> "observer" subjects' predictions of how they would rate the various factors were statistically identical to the actors' reported ratings. Hence, observers' ratings were also inaccurate (except for the effect of academic credentials), but no less accurate than those of the actors who actually experienced the various factors in judging Jill. Nisbett and Bellows interpret these results as evidence that the actors did not actually use the reasons they reported when they judged the candidates. Rather, they were unaware of how various factors influenced most of their judgments, and they retrospectively theorized about these influences in the same way the observers theorized about what factors *would* influence them. And in both cases, the theories were generally mistaken.[75]

Findings like these suggest that discrepancies can arise between the factors which *we think* are causes of our actions, and the factors which *really are* causes. Over the next four sections, I'll discuss some explanations of that fact.

3.8.2. Dissonance, confabulation, and rationalisation

Cognitive dissonance arises when there is some conflict between our actions and our attitudes, beliefs or desires. Research experiments show that we tend to revise our attitudes, beliefs and desires in order to resolve this conflict.[76] The effect is often so strong that we forget what our original attitudes, desires and beliefs were: we imagine instead that the attitudes we form *after* the action are the ones we had beforehand. For example, people asked to make a speech or write an essay defending a view they do not hold will often subsequently take up that view, and then (mistakenly) believe that it is the view they held before beginning the task.

The phenomenon of dissonance helps to explain why we fail to notice errors in our predictions about our own future actions. It also helps to explain why we may sometimes F-

[74] One of my five key cases is based on this experiment – see section 3.2.3. For a more detailed discussion of this experiment, see Nahmias 2001:210-214.
[75] Nahmias 2001:212.
[76] For more details of research on cognitive dissonance and confabulation, see for example Wegner 2002:171-186.

accept an action which we did not expect to perform. Consider Samantha in the 'Good Samaritan' example.[77] If we had asked, beforehand, about her attitude to someone walking past a slumped man when running late to give a talk, it's likely that she would have regarded the action as ethically wrong. However, if we ask her attitude *after* she herself has performed that action, she may well regard it as acceptable in the circumstances – because of a dissonance effect.

But at the same time, dissonance seems to pose a threat to my proposed CFA condition.

> *CFA* The agent would F-accept his action, if he knew all the proximate and relevant causes of the action.[78]

After her action, Samantha may F-accept it, even though she would not have done so beforehand. This seems to present an obstacle to determining whether the action is truly her own.

This problem can be overcome, I think, by stressing that the test of F-acceptance in the counterfactual scenario applies *at the time of acting*. We are interested in whether the agent would F-accept her action if she learned of its causes at the same time as she performs it. (Of course, this test cannot be carried out in practice; but since it is a counterfactual test, that does not matter.)

A possible experiment to *simulate* the counterfactual scenario might be hampered by dissonance. A delay between performing the action and learning of its causes might also allow a dissonance effect to occur in the agent. Therefore, it's often useful to compare agents' reports with the attitudes of people who have not performed the action themselves. We can examine our own reactions to each of the five key cases I've discussed. Imagining ourselves in those situations, most of us would expect not to F-accept our actions.

Another interesting psychological effect, apparently related to dissonance, is confabulation of explanations for acting.[79] An agent who lacks a plausible explanation for her own action will often confabulate one when asked – in other words, she will simply *invent* a suitable explanation. Furthermore, since confabulations are not deliberate, she will usually not realise that the confabulated explanation is invented. In many cases the confabulated explanation is also spurious (false). Thus Samantha, if asked why she did not stop to help the slumped man, might give as a reason that he looked threatening; she might believe this explanation, despite it being entirely false.

Working with people who had undergone commissurotomy (so-called 'split-brain' patients), Michael Gazzaniga suggested that confabulation is a function performed in the left hemisphere of the brain.[80] That it can occur in normal people in everyday situations has been shown in various experiments. A classic example is Nisbett and Wilson's work on the "position effect".[81] In one experiment, people in a shopping mall were asked to select their preferred pair of nylon stockings from four options on a table. The subjects were not told that the pairs were in fact all identical. The distribution of subjects' choices was heavily skewed toward the pair farthest to the right (this is the position effect). Asked to explain their choice, none of the subjects reported the position of the stockings as a factor. When position was suggested to them as a possible factor, almost all denied that it could have been. They pointed instead to superior colour, texture, quality of workmanship and so forth – all spurious reasons, since the stockings were identical. These are the sorts of explanations which we believe are

[77] See section 3.2.1.
[78] This is the simplified version of the condition – see section 3.5.
[79] It seems that we often confabulate *intentions* to act, as well as explanations for acting.
[80] For accounts of Gazzaniga's work and other research on commissurotomy patients, see for example Wegner 2001:181-184, Feinberg 2001:90-105 and Mackay 1987.
[81] For an account of these and other similar findings, see Nisbett and Wilson 1977, or Nahmias 2001:175-177 and 208-214.

important in our decision-making, and so – naturally enough – they are the sorts of explanations we give when asked to provide one. Experiments of this kind demonstrate that introspection is not always a reliable guide to our own reasons for acting.

Why do we not notice that we confabulate so readily? Part of the answer is that we confabulate only when we need to give some explanation for an action – if we are *asked* to give a reason, for example, or if we question ourselves about it. Many of our actions are simply not explained to anyone, and so there is no need for a spurious explanation, nor any opportunity to doubt it. Samantha might well not confabulate a reason if she is not questioned about her action. Meanwhile the usefulness of action-explanations does not depend only on their accuracy. One purpose of action-explanation is to influence the reactions of other people. Confabulation facilitates effective "interpersonal influence", and it may be that this gain outweighs the potential benefits from more accurate explanations of our actions.[82]

A similar and more familiar effect occurs *before* acting. Suppose Jon is offered a promotion to a new job based on the other side of the country. He decides to decline the offer, citing the reason that he does not want to move from his current home. However, his best friend suspects that the real reason why Jon declined the promotion was that he did not want to accept the extra responsibility and stress which that role would have entailed. A fortnight later, Jon accepts an offer of a similar but less demanding job – despite the fact that this job also requires that he move home. It seems that Jon's explanation for declining the first job was false. What may have happened, before Jon declined the first offer, is that he reached a plausible and acceptable – but false – explanation for his action. He *rationalised* an acceptable reason. Sometimes agents rationalise deliberately in order to mislead others about their real reasons for acting. Sometimes agents rationalise unintentionally in order to avoid thinking about uncomfortable reasons for acting, to justify an action which they feel uncertain about, or to justify an action which they want to perform despite a lack of acceptable reasons.[83]

Rationalisation, confabulation and dissonance are three parts of the explanation for our lack of awareness that our actions sometimes have subverting causes. The next part of the explanation is even less transparent to us when we act.

3.8.3. The hidden role of automatic processes

In section 3.6.2 I reviewed some specific examples of the influence on actions of unconscious automatic processes. I'll now discuss some of the broader implications of this research, specifically in so far as it suggests that our explanations of actions are inaccurate or incomplete.

In the early twentieth century, Freud proposed that unconscious motives, which the agent cannot report on, often generate actions. This central idea has remained influential, while many of Freud's suggestions about *how* and *why* the unconscious mind influences thoughts and actions have fallen out of favour. Freud famously compared the human mind to an iceberg: we are aware only of a small conscious part which lies above the surface, as it were, while a vast unconscious part lies hidden beneath.

In last few decades many important advances have been made in the study of "automatic" processes of which we are not consciously aware. Psychologist Timothy Wilson, who carried out some of the most influential studies in the field, summarises the modern view colourfully:

---

[82] See Baumeister 2008:76-78 for discussion of the relationship between action-explanations and social interaction.

[83] Haidt discusses the phenomenon of "motivated reasoning" in more detail (2006:63-66).

> When [Freud] said … that consciousness is the tip of the mental iceberg, he was short of the mark by quite a bit – it may be more the size of a snowball on top of that iceberg. The mind operates most efficiently by relegating a good deal of *high-level, sophisticated thinking* to the unconscious … The adaptive unconscious does an excellent job of sizing up the world, warning people of danger, setting goals, and initiating action in a sophisticated and efficient manner.[84]

John Bargh, another pioneer in this field, concludes that conscious thoughts are not causes of some actions:

> conscious acts of will are not necessary determinants of social judgment and behavior; neither are conscious processes necessary for the selection of complex goals to pursue, or for the guidance of those goals to completion. Goals and motivations can be triggered by the environment, without conscious choice or intention, then operate and run to completion entirely nonconsciously, guiding complex behavior in interaction with a changing and unpredictable environment, and producing outcomes identical to those that occur when the person is aware of having that goal.[85]

Some researchers have gone even further, and suggested that conscious thoughts may be entirely epiphenomenal – that they are *never* causes of actions. Psychologist Daniel Wegner recently created controversy by appearing to entertain that view.[86] Neuroscientist Benjamin Libet has also stirred a great deal of debate with an ingenious experiment from which he concludes that conscious intentions "to act now" occur late in the sequence of processes which lead to action – too late for them to be causally effective.[87]

No part of my argument requires the claim that conscious thoughts are always epiphenomenal.[88] My claims are that *subverting* causes are sometimes (or often) among the causes of actions.[89] These claims are compatible with the possibility that conscious thoughts are *always* among the causes of actions. It may be that *both* a conscious thought and a subverting cause are causes of an action – in the sense that both are necessary for the action. Or it may be that in some cases a conscious thought is a *sufficient* cause of an action, but the subverting cause *is a cause of that conscious thought*.[90] All of these possibilities are compatible with my claim that actions sometimes (or often) have subverting causes.

It seems to me that automatic processes *themselves* are often causes of actions. However, I don't need this claim, so I don't rely on it.[91] In section 3.3 I defined causes of actions as *proximate, relevant* events without which the action would not have occurred. An objector may take the view that automatic processes are part of the 'background' conditions, and that the *relevance* condition is not met, so that automatic processes cannot be causes of

---

[84] Wilson 2002:6-7 (my italics). The "adaptive unconscious" is Wilson's collective term for the huge number of unconscious processes which facilitate conscious thoughts and actions.

[85] Bargh 2005:52.

[86] See Wegner's response (2004b) to criticisms by Jack and Robbins (2004) and Kihlstrom (2004), among others. Mele (2009) carefully unpacks some of Wegner's statements which suggest that conscious thoughts may be epiphenomenal, before criticising them.

[87] See for example Libet 2002. For criticisms of Libet's philosophical conclusions, see for example Dennett 2003a chapter 8, or Mele 2009 chapter 4.

[88] In fact I think that claim is extremely unlikely to be correct.

[89] These are my claims (A) and (B) from section 3.6.

[90] In such a case the subverting cause is a cause of the action, if it is proximate to the action in the sense discussed in section 3.3.

[91] I think that some of those processes are *subverting* causes of actions, but I don't rely on that claim either. For a review of the research on automatic processes which has implications for responsibility and freedom, see for example Bargh 2005 and 2008, or Suhler and Churchland 2009.

actions. Even if that is conceded to the objector, though, there are three further points which I think are undeniable. First, automatic processes play an essential role in the causation of many actions. Even if we don't acknowledge them as causes in their own right, it is clear that automatic processes mitigate and facilitate the causation of actions by factors which we do consider to be causes. Second, agents are not aware of the operation of automatic processes. Third, agents do not understand the function of many automatic processes.

Bearing these points in mind, it is plausible that agents would sometimes be surprised to learn about the causes of their actions, even if we exclude automatic processes from consideration as causes. Since the operation of automatic processes is opaque to us, we cannot introspect the entire process by which any action is caused. Sometimes we know of a *possible* cause, and the action, and we incorrectly infer causation. Further insight into this inference is provided by the work of psychologist Daniel Wegner, which I'll discuss next.

3.8.4. Our experience of agency

According to Wegner, our experience of having acted – which I call the "experience of agency" – does not arise from introspective access to causal processes.[92] Instead, on Wegner's "theory of apparent mental causation", an agent's experience of agency is *inferred* from a correlation between his prior conscious thoughts about the action and his awareness of the action taking place.[93]

There are three key factors in the interpretation. The first is *priority*: a relevant conscious thought, such as an intention, occurs before the action. The second factor is *consistency* between the prior thought and the action. The third is *exclusivity*: the action is not accompanied by other potential causes.

I can't do justice to the evidence for Wegner's theory here, let alone defend it properly. But, broadly speaking, there are three sources of evidence for the theory. One is experiments in the psychology lab, where the priority, consistency and exclusivity factors can be carefully adjusted, and the subject's experience of agency is altered as a result. Another source of evidence is instances outside the laboratory where the experience of agency doesn't reflect actual causation of action. For example, Doug uses dowsing rods to locate underground water. He walks around holding two metal rods, and when the rods move he thinks that movement is caused by water under the ground. In fact Doug moves the rods himself by tilting them very slightly, but he has no experience of agency. On the other hand, Doug sometimes experiences agency when he is not causally effective – for example when he clutches his lucky charm to help his football team to score a goal.

The third source of evidence comes from the physical structure and evolution of the human brain. It seems likely that the ability to experience agency – which enables us to pick out our own actions and accept responsibility for them – is a relatively recent addition. The experience is most useful in animals which live in complex social groups, and can communicate the experience or its impact. Wegner suggests that most of our mammal ancestors had no use for such an experience. On the other hand, it's likely that we share with these ancestors much of the neural infrastructure which actually controls bodily movements. If a new cognitive faculty evolved to provide an experience of agency, the most likely way for it to work is by interpreting correlations between experiences which already exist – in this case the experience of considering a possible action, and the experience of witnessing its

[92] See Wegner 2001 chapter 3. Wegner calls this "the experience of conscious will" – which he describes as "experienced when we perform an action … [a] feeling of voluntariness or doing a thing "on purpose"" (Wegner 2001:3).

[93] Here I am ignoring the controversy about whether conscious thoughts are ever causes of actions (and about whether Wegner believes that they are). See Mele 2009 for a thorough discussion of that issue. I am focusing here on Wegner's claims about the source of our experience of agency.

performance. This is much more likely to have evolved than the alternative: a new faculty which is physically connected to the 'older' parts of the brain's architecture – the parts which trigger and control bodily movements – in order to monitor their activity directly.

If Wegner is right, agents do not know *from introspection* that their own actions are caused by conscious mental states or events. Instead, they *infer* that causation. It follows that an agent does not know from introspection that any particular conscious mental state or event was a cause of a particular action. The agent does not know from introspection that any particular conscious thought – whether an intention, decision, desire or belief – was a cause of his action.[94] When the agent concludes that one of these was a cause of the action, he does so by an inference based on the priority, consistency and exclusivity factors. This inference is fallible, as Wegner demonstrates experimentally.

I think this theory, if correct, helps to explain why we may sometimes be surprised to learn of the causes of our actions. Take the case of Ingrid, who interrupted a conversation after completing the rudeness-themed puzzle (see section 3.2.4). Suppose that, just before acting, Ingrid has a conscious thought that she needs to leave soon to meet someone. This thought is prior to and consistent with her action of interrupting the conversation (she can say what she wants to say and then leave for the meeting). As far as she is aware, the thought is also an exclusive explanation: she does not realise that the content of the word puzzle might be another possible explanation. She concludes that her thought about the meeting is a cause of her interruption. According to Wegner's theory, she cannot know from introspection that this is true. The thought about the meeting might be entirely coincidental, and not a cause of her action. Instead, and unbeknownst to her, working on the priming puzzle *is* a cause of her action.

3.8.5. Errors in interpretation and assessment of situations

Situational factors can have surprising effects on actions, as I discussed in section 3.6.1. Perhaps equally surprising is that they sometimes operate by distorting the agent's assessment of the situation in which she acts.

The 'Good Samaritan' and 'market research' experiments are among the most disturbing in the situationist literature. In one, 90% of subjects walked past a slumped man after being told they were running late to give a presentation.[95] In the other, 93% of subjects continued to fill out a market research questionnaire despite the researcher's cries of distress from the next room, if someone was sitting passively next to them.[96] The most natural explanation of these findings is very alarming: it seems as though the great majority of subjects callously failed to help someone in obvious need of assistance, and knowingly disregarded a clear and important moral demand.

However, the psychologists leading the studies found that this natural explanation is too simplistic. Instead, Latané and Darley devised a model of "the intervention process", which includes the steps through which "bystanders" (such as the subjects in the experiments above) may pass when they do intervene to help someone in need.[97] The three steps are these: the bystander "must *notice* that something is happening, *interpret* that event as an emergency, and decide that he has *personal responsibility* for coping with it".[98] All three of these steps may be influenced by the situational factor, with the result that the subject does not help.

---

[94] Wegner himself does not draw this conclusion explicitly; his objectives are different from mine. But these points seem to me to follow from Wegner's stated conclusions.

[95] Darley and Batson 1973. See section 3.2.1.

[96] Latané and Rodin 1969. See section 3.2.2.

[97] Latané and Darley 1970, chapters 4, 10 and 13. The model is based on a series of experiments on the theme of "bystander intervention", like the two experiments I mentioned above.

[98] Latané and Darley 1970:121.

Take for example the 'market research' experiment. At the first step, the situational factor (the presence of another person filling out a questionnaire) might lead subjects not to *notice* the cries of the researcher, or to notice them more slowly. But this effect was found not to be very pronounced.[99]

The influence of the situational factor at the second step, however, was much more pronounced and very interesting. Many subjects appear not to have *interpreted* the situation as an emergency:

> They did not act because they thought there was no emergency, or that the emergency was not serious, or that it would be inappropriate to act … We think they were correct when they stated that if their interpretations had been different, they would have intervened.[100]

This suggests that, in many cases, the situational factor influenced the subject's action *by influencing his interpretation of the situation which he faced*.

At the third step, too, the influence of the situational factor can be significant. In particular, when there are more than one witness to an event, the "onus of responsibility is diffused", and each individual bystander is less likely to intervene.[101] This may help to explain why many subjects did not respond to the market researcher's cries: they perceived a lesser moral onus to help, compared to subjects in the control experiment, because there was a second witness present.

The full version of Latané and Darley's model is more complex still. Subjects do not necessarily move sequentially through the three stages before helping or not helping. Instead three further phenomena, which they call *cycling*, *blocking*, and *commitment*, may hinder helping.[102] When "cycling" occurs, the bystander's *interpretation* of the event is influenced by the onus of *responsibility* which he feels. For example, a subject hearing the market researcher's cries might initially interpret the situation as an emergency, but then feel only a small onus of responsibility to help because of the presence of the second witness. This in turn may lead him to *reinterpret* the situation as much less serious. When "blocking" occurs, a subject does not come to any decision about what to do. Rather than deciding not to help, for example, confusion and conflict leads him simply not to reach a decision about whether to help. The third phenomenon, *commitment*, can follow blocking. A lack of a decision to help effectively commits the subject to the alternative path of *not* helping, as it were by default.

We should not assume that subjects are consciously aware of the influence of the situational factor on these interpretations and decision-making processes.[103] Over a series of experiments where the situational factor was the presence of a bystander, the researchers found the following.

> We asked this question every way we knew how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present. This denial occurred in the face of results showing that the presence of others did inhibit helping … People seem quite sincerely and genuinely unaware of the various ways in which they are influenced in their definitions of physical and social realities by the behavior of other people.[104]

---

[99] Latané and Darley 1970:87.
[100] Latané and Darley 1970:89.
[101] Latané and Darley 1970:90.
[102] Latané and Darley 1970 chapter 13.
[103] It may well be that many important parts of these processes occur unconsciously – see section 3.8.3 above.
[104] Latané and Darley 1970:124. Of course, it may be that the presence of bystanders was not a cause of action for *some* of the subjects. But the fact that almost all the subjects denied that it was a cause shows that many of them were mistaken.

If Latané and Darley's model is at all accurate, then it seems that the "natural explanation" of the experimental findings is deeply flawed. Subjects did not, in the main, callously fail to help people in obvious need of assistance, and knowingly disregard clear and important moral demands. Instead, the situational factor often influences subjects' interpretations of the situation and their assessments of the moral onus upon them to act.

The experiments covered by Latané and Darley's model are some of the most shocking in the situationist literature. It seems likely that similar phenomena occur when other kinds of situational factors are causes of actions.

3.8.6. Conclusions from the experimental evidence

The experimental evidence I have reviewed over the past several sections establishes, I think, that the following two claims are extremely plausible:

> (A1) The five examples of subverting causes in section 3.2 are not isolated incidents which occur because of dubious experimental methods. Rather, they are representative of a large body of credible experimental evidence in mainstream fields of psychology.
>
> (B1) There are good grounds to believe, on the basis of existing experimental evidence, that many everyday actions have subverting causes.

I've given overviews of the research fields in which the effects in my five key cases were discovered. The actions in those five key cases are just like many everyday actions. Many other studies have revealed similar effects. Many of these have studied real-life everyday actions, or actions that are just like everyday actions. I also rejected some objections to these claims. Nevertheless, it is difficult to believe that everyday actions can have subverting causes. I discussed evidence which suggests explanations for our failing to notice that everyday actions have subverting causes.

I will now present a brief inductive argument in support of my third claim:

> (B2) There are good grounds to believe that there are many more subverting causes of everyday actions as yet undiscovered by experimenters.

Most of the evidence which supports claims (A1) and (B1) has only recently been discovered. Almost all of the evidence dates from the last fifty years, and much of the most interesting evidence has come in the last decade. A lot of the evidence has been extremely surprising: we did not expect to discover that our actions can be influenced by the factors which have been pinpointed. Some of the research fields from which the evidence has come are still in their infancy, and the experimental methods used are steadily becoming more and more sophisticated. Therefore it is very likely that many more subverting causes and effects will be discovered by experimenters in the coming years. This supports my claim (B2).

Claim (A1) above supports the following claim:

> (A) There are actions with subverting causes.

This in turn supports a key premise in my "main argument" against Frankfurt in section 3.4: the actions in the five key cases have subverting causes. Claims (B1) and (B2) support the following claim, which will be a premise in my "sceptical argument" in chapter 7:

> (B) There are many everyday actions with subverting causes.

The term "many" in my claim (B) is deliberately loose, and indicates a range of possibilities. Someone who finds it plausible that there are *very* many actions with subverting causes will find that my argument in chapter 7 leads to a very sceptical conclusion.

3.9. CONCLUSIONS FROM THIS CHAPTER

I'll now recap my conclusions from the chapter as a whole. I've argued that:

(A) There are actions with subverting causes.[105]

I've given examples of such actions, discussed evidence for the claim that there are such actions, dealt with some important objections to that claim, and explained why it is difficult to accept.

In my "main argument", I argued that F-acceptance of an effective desire is not sufficient for its being truly the agent's own. In some actions with subverting causes, Frankfurt's identification condition (F-acceptance) is met, and yet the effective desire is not truly the agent's own. From this, it follows that F-acceptance of an effective desire is not sufficient for direct moral responsibility for the action performed.[106] I discussed five key cases of actions with subverting causes, for which this argument can be run. A single counterexample is enough to show that F-acceptance is not sufficient for direct moral responsibility.

I also made a positive proposal: that my CFA condition is necessary for direct moral responsibility. An agent's effective desire is truly her own only if the agent *would F-accept it*, if she knew the causes of the action.

I argued that followers of Frankfurt should accept these conclusions, since my arguments employ the central elements from his own theory in ways which are consistent with the theory's objectives.

Finally, I argued for the following claim:

(B) There are many everyday actions with subverting causes.

I will draw on claim (B) in my "sceptical argument" in chapter 7 below.

[105] An action has subverting causes when it's true that: if the agent knew all the causes of her action, she would not F-accept the action as her own.

[106] This conclusion follows if we accept (as Frankfurt does) the following identificationist axiom: For direct moral responsibility for an action, the effective desire must be truly the agent's own. This is statement (4) in section 2.2.

4. OBJECTIONS TO MY CONCLUSIONS

In chapter 3 I set out my main argument against Frankfurt's theory of moral responsibility. I concluded that Frankfurt's condition of identification, which I labelled "F-acceptance" of the effective desire, is not sufficient for direct moral responsibility.[1] I argued that when an action has subverting causes, the effective desire which moves the agent to act is not truly her own.[2] On an axiom of identification which Frankfurt himself holds, an agent cannot be directly morally responsible for an action if the effective desire is not truly her own.[3]

I also made a positive proposal – namely that the following condition is necessary for direct moral responsibility:

> *CFA* The agent would F-accept the effective desire, at the time of acting, if he knew, at the time of acting, all the proximate and relevant causes of that desire's being effective.[4]

In this chapter I'll discuss some objections to my main argument and to my positive proposal. The first two of these objections involve cases which seem to create problems for my proposed CFA condition. In section 4.1 I'll discuss actions with situationist causes which do *not* subvert moral responsibility, on my account. These actions seem in many respects very similar to the five key cases which I employed in my main argument: it could be objected that the threat to moral responsibility is even greater than the one I've portrayed. I'll argue that compatibilists can give plausible replies to this objection.

In section 4.2 I'll examine cases of regrettable actions, in which it is obvious that the agent is DMR, and yet it appears that my CFA condition is not met. These would be counterexamples to my claim that the CFA condition is necessary for DMR. However, I'll show that the CFA condition is indeed met in these cases, and so the objection dissolves.

Then, in section 4.3, I'll discuss several objections which have a common structure. Agents in the five key cases of actions with subverting causes, who do not meet the CFA condition, might nevertheless seem to be directly morally responsible for their actions. This would undermine both of my arguments from chapter 3. This objection can be presented in several forms. Against each form, I'll defend my view that the agents are not DMR for these actions.

Some of these objections would not be raised by followers of Frankfurt. By considering such objections, I am perhaps going beyond what I need to do to defend my arguments against Frankfurt's position. However, I think it's still useful to consider all of the objections I'll discuss, because (if I can answer them successfully) they help to strengthen the conclusions I reach.

4.1. NON-SUBVERTING SITUATIONIST CAUSES

One important source of evidence about subverting causes has been research in situationist psychology, as I discussed in section 3.6.1. However, the surprising effects discovered by situationist research are not always subverting. I've already noted that certain agents in the same situations as my five key cases might F-accept their actions after learning of their causes. Those causes, although unknown, would not be subverting. Furthermore, in some experiments it seems likely that the surprising situationist cause is not subverting for the *majority* of subjects.

---

[1] An agent F-accepts his effective desire if he accepts it as his own, and is satisfied with that acceptance. Being satisfied is a matter of having no interest in making changes. See section 2.1.3.

[2] An action has subverting causes when it is true that: if the agent knew all the causes of her action, she would not F-accept the action. (This is a simplified definition; for the full version see section 3.1).

[3] This axiom is statement (4) in the set which define Frankfurt's position in section 2.2.

[4] This again is a simplified version of the condition; for the full definition see section 3.5.

For example, in a famous situationist experiment, people exiting a phone booth in a busy shopping centre (who were the unsuspecting subjects) encountered a researcher who dropped a folder of papers in their paths.[5] Some 88% of subjects who had just found a dime in the coin return slot of the telephone (placed there by the experimenters) stopped to help pick up the scattered papers. But of those subjects who found no dime, only 4% stopped to help. The experimenters speculated that subjects are much more likely to help after finding the dime because they are in a better mood. Further research suggests that apparently small stimuli which influence moods can produce significant behavioural effects.[6]

It's possible that the "mood effect" in cases like this one operates (at least for some subjects) by heightening the person's awareness that they are confronted with a moral demand. That is, without the mood effect provided by finding the dime, perhaps some people do not notice the moral demand upon them to help the researcher to pick up her papers.[7]

It's clear that finding a dime in the return slot was a cause of the helping action in very many cases. It seems likely that the majority of these subjects, if told of the causes, would F-accept the helping action they performed. Meanwhile, of those subjects who did *not* find a dime, it seems very likely that the large majority did F-accept their action when walking past the researcher and her scattered papers, and that they would still do so if they knew all of their action's causes. These subjects already do know as much about the causes of their action as the experimenters know: I'm assuming that we wouldn't include the *absence* of a dime in the slot as a cause of their action.

There is something paradoxical about this kind of case, which I think can give rise to four objections to my proposal. All four objections lead to a more sceptical conclusion than mine about the prevalence of moral responsibility. Since my position is already sceptical, I'm not troubled by such objections. Nevertheless in each case I think these objections can plausibly be rejected by compatibilists.

In the first objection, consider an agent – I'll call him Walter – who does not find a dime, and does not help pick up the scattered papers but walks away instead. Suppose that Walter F-accepts his action, and that the following counterfactual (X) is true of him: if he knew all the causes of his action, he would still F-accept it.[8] Walter meets the CFA condition which (I proposed) is necessary for DMR for his action.

However, suppose the following counterfactual (Y) is also true of Walter: if he had found a dime in the slot, he would have helped to pick up the researcher's papers. Suppose further that counterfactual (Z) is true as well: if Walter learned that counterfactual (Y) is true, he would not F-accept the action as his own. The objector claims that this action is not truly Walter's own. It follows that Walter is not directly morally responsible for it – if being truly one's own is necessary for DMR for an action.[9] In the objector's view, what subverts Walter's moral responsibility here is not a cause of his action, but rather what we might call *an enabling condition* of his action.[10] In this case, the enabling condition is *that there was no dime in the slot in the phone booth*.[11]

---

[5] Isen and Levin (1972). For more discussion of the experiment, see Doris 2002:30 or Nelkin 2005:185.
[6] See Doris 2002:30.
[7] See Doris 2002:180n5.
[8] The *absence* of a dime in the slot is not a cause of his action.
[9] This identificationist axiom is number (4) in the list of statements defining Frankfurt's position, in section 2.2.
[10] Albeit that this is an enabling condition of a rather strange sort – a *negative* enabling condition.
[11] Note that an enabling condition which would have led the agent to act otherwise than he did is not necessarily subverting, on the objector's view. For example, suppose this counterfactual (R) is true: if Walter had won the lottery yesterday, he would not have driven to the shopping centre today. This does not entail that the enabling condition *that Walter did not win the lottery* subverts his moral responsibility for his action of driving to the shopping centre. If Walter learned that counterfactual (R) is true, he would nevertheless still F-accept his actual action. So the enabling condition *that Walter did not win the lottery* is not subverting here.

Is it true that subverting enabling conditions prevent (direct) moral responsibility? I find myself in two minds on this question. On the one hand, it seems implausible to say that Walter's moral responsibility for an action can be influenced by *the absence* of something which would have been a cause of his performing a different action. We are accustomed to excusing agents from moral responsibility because of the presence of a cause of their action – for example an addiction or a compulsion. But we are not accustomed to excusing agents because of the absence of a cause of an alternative action. On the other hand, I am also drawn to agree with the objector, on the grounds that there is no obvious principled reason why a cause can excuse but an enabling condition cannot.

This objection leads to a more sceptical view than mine about the frequency of direct moral responsibility for actions. If enabling conditions can be subverting, then there are fewer actions for which agents are directly morally responsible. I assume that most compatibilists will resist this conclusion. My arguments are mainly directed toward compatibilists. Therefore I will assume that the objection fails, thereby coming down on the compatibilist side of the fence — since this gives no support to my sceptical argument in chapter 7. There I will argue that the prevalence of subverting causes means that there are many everyday actions for which we mistakenly hold agents morally responsible. That is already a strongly sceptical conclusion. Its impact is more powerful if it goes through without the support of the point made by the current objector.

The three remaining objections concern an agent (I'll call him Henry) who helps the researcher immediately after finding a dime in the return slot, and when the finding was a cause of the helping action. That a situational factor like this can be a cause of this action is certainly surprising. But suppose that Henry would F-accept his action, if he knew all of its causes. On my account, this cause is not subverting. The remaining objections make a case, in differing ways, that Henry is not directly morally responsible for his action. If that is correct, then an extra necessary condition is required for DMR, in addition to counterfactual F-acceptance.[12] Again this is a more sceptical conclusion than mine, since it leaves fewer actions for which agents can be DMR. But I think that compatibilists can resist these objections.

The second objector holds that Henry must know the most important causes of his action, in order to be directly morally responsible for it. The relevant sense of "important" will be difficult to define, but I'll set that point aside for the sake of argument.[13] To this objector I reply that her position is implausibly restrictive. It's not plausible to require that an agent must know all the most important causes of his action, in order to be DMR for it. I'm not aware of any philosopher who holds this objector's view.

The third objection might be made by an identificationist: she might argue that Henry's action is not truly his own in the way required for direct moral responsibility. Some condition of identification must be set out, which is not met in Henry's action. But I cannot think of an existing identificationist account on which Henry's action would be held to be not truly his own.

Instead, a fourth objection might be made. This objector claims that Henry's *authorship* of the action is undermined by the situational cause – that the action is *not really Henry's*, in the appropriate sense for direct moral responsibility. The objector must provide a definition of "authorship". And if this objection is separate from the third objection above, the required

[12] If it were true that Henry meets my CFA condition while not being DMR for his action, that would not refute any of my claims about CFA. In chapter 3 I argued that Frankfurt's identification condition is not sufficient for DMR, and that my CFA condition is *necessary* (though not sufficient) for DMR.

[13] A more stringent version of this objection would require that the agent must know *all* the (proximate and relevant) causes of his action. This would lead to an even more sceptical conclusion about the possibility of moral responsibility. As a result I think it is even less plausible than the objection I discuss here.

definition of authorship will not involve identification. Again I find myself in two minds about this objection. There does seem to be a sense in which Henry's authorship is impaired by the situational cause. But at the same time, it may be that the "mood boost" helps focus Henry's attention on the moral demand to help the researcher, so that he is, as it were, *more actively engaged in his action* of helping the researcher than he might have been if he had walked past her without the presence of the dime. This is unclear. Once again I'll take the line which gives least help to my sceptical argument in chapter 7. A restrictive definition of authorship would lead to a more sceptical conclusion about the prevalence of direct moral responsibility. I think it's unlikely that compatibilists would press this objection, so I'll assume for the sake of argument that it fails.

I conclude from this section that my main argument (in chapter 3) can be defended against various objections which draw on non-subverting situationist causes. These objections lead to conclusions about the prevalence of moral responsibility which are more sceptical than mine. I am not troubled by the prospect that my position could be outflanked in this way. At the same time, I think these objections can plausibly be rejected by compatibilists.

4.2. REGRETTABLE ACTIONS WITHOUT SUBVERTING CAUSES

Another type of problem case for my argument involves regrettable actions which do not have subverting causes. In some such actions, it may seem that my CFA condition is not met, and yet the agent is DMR. This would refute my conclusion, in chapter 3, that the CFA condition is necessary for DMR, and for the action being truly one's own.

Consider Philip, who hears a fire alarm in his hotel room and sets off down the main stairs. Unbeknownst to him, there is a fire on the main staircase, which is what triggered the smoke detectors and caused the fire alarm to sound. When he sees the smoke, Philip has to retrace his steps and take the external fire escape instead to reach safety. If, at the time of setting off, Philip learned all the causes of this action, he might learn that the fire which is a cause of his action *is in the main staircase*. Now if he gained that knowledge, Philip would not F-accept his action, because he would immediately reverse it and take the external fire escape instead. It may seem that Philip's action is not truly his own on my account, and that he cannot be DMR for walking down the main stairs.[14]

That would be an extremely implausible conclusion; but in fact it does not follow from the full version of my CFA condition, given in my original presentation in section 3.5:

> *CFA* The agent would F-accept the effective desire, at the time of acting, if he knew, at the time of acting, all the proximate and relevant causes of that desire's being effective (*except in so far as any additional knowledge gained gives him new information about means to satisfy his desires*).

Among Philip's desires, in the actual situation, are the following: a desire to go down the main staircase (the effective desire to act), a desire to escape from a possible fire, and a desire to survive the night and continue living. If he learned that there was a fire, *and that the fire was in the main staircase*, Philip would have new information about the means to satisfy his desires. This information is not available to him in the counterfactual scenario which is relevant for determining whether his actual action is truly his own. In that relevant counterfactual scenario, he would learn *that a fire was a cause of his action*. He would not learn *that the fire is in the main staircase*. Hence he would continue to F-accept his effective

[14] The conclusion about DMR follows if being truly one's own is necessary for DMR – this is axiom (4) in the set defining Frankfurt's view (see section 2.2).

desire and the action which results from it. The CFA condition is therefore met, and Philip can be DMR for his action.

Consider a similar example involving a misunderstanding. Simon is swimming in the sea and sees a flag flying above the lifeguard station. He swims back to shore, thinking that the flag indicates danger. But in fact the flag signifies that conditions are safe for swimming. A cause of Simon's action is the senior lifeguard assessing the situation as safe and hoisting the flag. Simon's desires include the desire to swim back to shore (the effective desire to act), a desire to avoid swimming in dangerous waters, and a desire to survive his swim and continue living. If he learned that the lifeguard had raised the flag, and *that the flag signifies that conditions are safe*, Simon would have new information about the means to satisfy his desires. This information is not available to him in the counterfactual scenario which is relevant for determining whether his actual action is truly his own. In that relevant counterfactual scenario, he would learn only *that the raised flag was a cause of his action*. Hence he would continue to F-accept his effective desire and the action which results from it. The CFA condition is met, and Simon can be DMR for his action.

Someone might object that the final italicised clause in the definition of the CFA condition represents an unprincipled and wholly ad hoc reaction to a potential problem with my proposal. This problem is posed by regrettable actions in which some key information about the causes of his actions is unavailable to the agent. In response to this objection I appeal to a clear difference between the cases of Philip and Simon on the one hand, and those of Samantha and Martha on the other.[15] Samantha walks past a slumped man because she is running late. Martha continues to complete a questionnaire despite a researcher's cries in the next room. Knowing of the causes of their actions, and understanding the nature of situationist causes, it is easy to judge that these agents' decision-making processes are subverted, so that the effective desires are not truly their own, in the sense which Frankfurt aims to capture.

The same is not true of Philip and Simon. We might say that their decision-making processes are constrained by a lack of information, but that lack is not subverting. Their effective desires are truly their own, in the Frankfurtian sense.

I conclude that "regrettable" actions such as these do not undermine my conclusions.

4.3. DIRECT MORAL RESPONSIBILITY DESPITE SUBVERTING CAUSES

In this section I'll address several forms of an objection in which the following claim is made: the agents in my five key cases are DMR for their actions, despite those actions having subverting causes. This claim, if true, would affect two of my arguments, which I'll discuss in turn.

First, the objector's claim would affect my main argument against Frankfurt, in section 3.4. In that argument I made use of the following statements which partly define Frankfurt's position:[16]

(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).[17]
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility (A→R).

[15] Samantha and Martha's cases appear in sections 3.2.1 and 3.2.2.
[16] See section 2.2 for a discussion of these statements.
[17] An agent F-accepts her action if she accepts it as her own, and is satisfied with that acceptance. Being satisfied is a matter of having no interest in making changes.

I argued that statement (5) is false, since agents in the five key cases do F-accept their effective desires, and yet those effective desires are not truly their own. If that is correct, then (6) is false if (4) is true.[18] It also follows that the agents in the five key cases are not directly morally responsible for their actions, since their effective desires are not truly their own, and – according to (4) – the desires' being truly their own is necessary for DMR.

Now, if it turns out (as will be claimed in the objections in this section) that the agents in the five key cases *are* directly morally responsible for their actions, then one of two conclusions must follow. It could be, on the one hand, that axiom (4) is false. Then it would be possible that the agents' effective desires are not truly their own (as I argued) and yet they are DMR for their actions (as the objector argues). Or, on the other hand, it could be that axiom (4) is true and so it must follow that my claim is false: the agents' effective desires *are* truly their own.

Thus the objections I'll consider in this section do not necessarily establish that my claim about the five key cases (that that the agents' effective desires are not truly their own) is false. It could be, instead, that axiom (4) is false. But having noted that point I will set it aside. My main objective is to show that Frankfurt's statement (6) is false, *even if (4) is true*, so I will continue to assume for the sake of that argument that (4) *is* true.

It's also important to remember that, for my main argument, I need only one example of an action in which the agent's effective desire is not truly her own (and hence she is not DMR for her action). In order to refute this argument, the objector would need to show that *none* of the agents in the five key cases are DMR for their actions.

Meanwhile, the objector's claim, if true, would also affect my positive proposal that the CFA condition is necessary for an agent's effective desire being truly her own.[19] From that it follows – if axiom (4) is true – that the CFA condition is necessary for the agent's direct moral responsibility for the action.

In the five key cases, the agents do not meet the CFA condition: they would not F-accept their effective desires if they knew all the causes of their actions. If the objections in this section succeed, and the agents are DMR for their actions in the five key cases, then one of two conclusions must follow. It could be that axiom (4) is false. But, as above, I will set aside that possibility. The other conclusion is that the CFA condition is not necessary for DMR. In order to refute my positive proposal, the objector needs only to show that there is a *single* case in which the agent is DMR but does not meet the CFA condition.

I won't consider the objector's claim for every one of the five key cases. But it is my aim to defend my positive proposal, so I hope to show how the objection can be refuted for *all five* of the key cases. Meanwhile, refuting the objection for *at least one* of the key cases will suffice to defend my main argument above.

I'll discuss the following forms of the objection. In section 4.3.6 I'll consider the claim that agents of actions with subverting causes can be DMR for performing them on the grounds that they *feel* morally responsible for doing so. But I'll begin by addressing the important claim that agents in the five key cases are DMR for their actions because *they are blameworthy for performing them*, despite the actions having subverting causes. I'll introduce this objection in section 4.3.1, and then, over the four subsequent sections, discuss four forms in which it can be advanced.

[18] If ¬(A→O) while (R→O), then it follows that ¬(A→R).
[19] See section 3.5.

4.3.1. The agent is blameworthy for the action

The reasoning shared across this family of objections is as follows. The objector claims that, in at least one of my five key cases, the agent is blameworthy for performing the action. If she is blameworthy, she must be morally responsible for the action – since moral responsibility is necessary for blameworthiness.[20]

(It's conspicuous that this reasoning employs morally bad actions, for which the agent is considered blameworthy. A parallel form of reasoning could be advanced for agents who are *praiseworthy* for actions which have subverting causes: they must be morally responsible for those actions. However, I've never encountered this form of the objection.[21] Perhaps we are more strongly motivated to insist that agents are blameworthy for bad actions than that agents are praiseworthy for good actions.[22])

The point I make initially, in reply to the objection, is that blameworthiness does not imply *direct* moral responsibility. An agent may be blameworthy for an action because she is TMR for it. In fact, I think that some of the intuitive pull toward blaming agents in the cases I discuss is attributable to the possibility that they are TMR, even if not DMR. Our intuitions do not distinguish very clearly between TMR and DMR. In order to keep the distinction clear I'll use the phrase "directly blameworthy" to mean "blameworthy in virtue of the agent's being DMR – as opposed to TMR – for her action".

I have two further replies to the objection. The first is simply that this objection would probably not be made by an identificationist like Frankfurt.[23] It's an identificationist axiom that direct moral responsibility is possible only if the action is truly the agent's own.[24] I've argued that, when an action has subverting causes, it is not truly the agent's own.[25] If my argument there is persuasive (as I argued that it should be to followers of Frankfurt), then the present objection is not open to anyone who holds that identificationist axiom.[26] However, I won't rely only on this first reply.

My second reply is to deny the objector's claim that the agent is directly blameworthy in one of my five example cases. The objector can make this claim in several ways, which I'll discuss over the next few sections (4.3.2 to 4.3.5). I won't consider all five key cases of actions with subverting causes in each form of the objection. For simplicity I'll focus on one of those cases – Martha's completing her questionnaire despite the researcher's cries.[27] I've picked this case because it's the most difficult for me: it seems that Martha faces a substantial moral demand to help, in a situation of very obvious need, and several ways in which an objector might plausibly claim that she is directly blameworthy for her action. By defending my conclusion that Martha is not DMR for her action, I will go a long way toward establishing how the conclusion can be defended in the other cases.

[20] I assumed that this principle is correct in section 1.1.3.

[21] Doris and Nahmias, too, when discussing the implications of situationist research, focus almost entirely on the possible implication that agents should be exculpated for morally bad actions, rather than the possibility that agents are not praiseworthy for good actions. (See chapter 5.)

[22] This may be one explanation for some intriguing recent discoveries in the new field of "experimental philosophy" – see for example Knobe 2006:138-139.

[23] In section 6.7.2 I'll argue that this objection would not be made by Fischer and Ravizza either.

[24] This was statement (4) in my summary of Frankfurt's position in section 2.2.

[25] See section 3.4.2.

[26] If the action is not truly the agent's own, then she cannot be DMR for it, and so she cannot be directly blameworthy for it.

[27] See section 3.2.2.

4.3.2. The agent allowed herself to perform the wrong action

In this first form of the objection, the claim I must refute is that Martha is blameworthy for her action because *she allowed herself* to complete her questionnaire rather than help the researcher.

This objector's explanation for Martha's action is something like this. Martha must have realised that the researcher was in need of help. She also realised there was some disadvantage in going to offer help, such as the risk of embarrassing herself in front of the other bystander who continued to fill out his questionnaire. She must have assessed the moral demand to help as greater than the moral demand to complete the questionnaire, but she allowed the disadvantage to weigh more heavily in her deliberation. She is directly blameworthy and DMR because she *allowed herself* to perform the morally worse action.

My reply is that Martha's case need not fit this description. As I discussed in section 3.8.5, situational factors can influence an action *by influencing the agent's interpretation* of the situation. Consequently, the objector's explanation for the action is not the only one. There are three possible explanations, which I'll discuss in turn.

The first explanation is as follows. Martha is not aware of the situationist research which reveals that the presence of a bystander can have a significant influence on an agent's interpretation of a situation. As a result of that influence, Martha misconstrues the situation as one in which the researcher's need is less morally pressing than her duty to complete the questionnaire.[28] Martha does not recognise that the greater moral onus on her is to help the researcher, and instead she continues to complete the questionnaire.

In this first case, I think it is unreasonable to hold Martha directly blameworthy for her action, or for failing to recognise that the greater moral onus on her is to help. She misconstrues the situation, but she does not *culpably* misconstrue it. Hence we cannot conclude that she is DMR on the grounds that she is directly blameworthy.[29]

The second possible explanation for Martha's action is this. Martha *has* read about situationist research which reveals that bystander presence can have a significant influence on an agent's interpretation of a situation. However, when she hears the researcher's cries she does not think about the information she has read. Perhaps she has forgotten about it, or perhaps it simply does not come to mind at the crucial time. And because of the bystander's presence, she does not recognise that the greater moral onus on her is to help the researcher; instead she continues to complete the questionnaire.

In this second case, depending on the circumstances, I think it might be reasonable to say that Martha culpably misconstrues the situation. We might say that, after reading about situationist influences on actions, Martha should have taken some steps to ensure that she recognised future situations in which those influences might be present. She might be blameworthy for omitting to take those steps. Whether she is in fact blameworthy will depend on such factors as whether there *were* any steps which she could have taken to prevent the later action, and whether she could reasonably have known what those steps were.[30]

But even assuming that Martha is blameworthy for not taking those steps, this establishes only that she bears *traced* moral responsibility for her action. Her moral responsibility for her current action is anchored in facts about her earlier omission, rather than in facts about the current action. She is moved to her current action by an effective desire

[28] In fact, it may be that Martha does not assess the researcher as being in need of help *at all*. (See section 3.8.5.)

[29] In this first case I think we should conclude that Martha is not TMR for her action either. But I do not need this point for my argument here.

[30] I think it's far from obvious that there was any culpable omission on Martha's part here. Most of us, upon learning of situationist research, can do little more to modify our future behaviour than to "try to pay attention to and overcome group effects when confronted with an emergency" (Nahmias 2001:189). However, I don't need this point for my argument.

which she would not F-accept, if she knew all the causes of her action. As I argued in section 3.4.2, this action is not truly her own, and so she is not *directly* morally responsible for it.

The objector might press the view that Martha is *directly* blameworthy for her action, on the grounds that she should have made use of the information about situationism that she had read – even if she did not recall that information at the time of acting. I think this is unreasonable. If Martha genuinely would not F-accept her effective desire on learning of the action's causes, then the action is not truly her own in the sense emphasised by Frankfurt as necessary for DMR. At any rate, I don't think that followers of Frankfurt would press this point, and it is to them that my main argument is addressed.

The third possible explanation for Martha's action is as follows. Martha does recognise the greater moral onus on her to help the researcher, but allows her desire not to look foolish in front of the bystander to prevail. In this case, the presence of the bystander is *not* a subverting cause. Martha knows that it is a cause of her action.[31] She has weighed the options, and decided to complete the questionnaire. She F-accepts this action, and would F-accept it if she knew all of its causes.[32] On my account, she *is* directly morally responsible for her action.

In summary, the objector's claim that Martha is directly blameworthy for her action is plausible if we assume that the third explanation for her action is correct. But on my argument, Martha *is* DMR for her action if the third explanation is correct. So this case does not present an objection to my argument.

But the third explanation is not the only possible explanation. On the other two explanations I've considered, Martha is not directly blameworthy for her action. Hence the objection fails in cases like these.

I think it's extremely implausible to think that the third explanation applies to every subject in the 'market research' study. That is, it's extremely implausible to think that everyone who completed the questionnaire rather than help did so in spite of recognising the moral onus to help as the greater of the two.[33] It's much more plausible that the first explanation applied to many subjects in the study.[34] Many agents in the same position as Martha would continue to complete their questionnaires, and not be DMR for their actions. This supports the conclusions I drew in chapter 3 about the five key cases including Martha's.[35]

So this objection fails to establish that Martha is blameworthy for her action because she allowed herself to complete her questionnaire rather than help the researcher.

4.3.3. The agent should have tried harder to resist the influence

Another way for the objector to claim that agents can be directly blameworthy for actions with subverting causes is as follows.[36] The objector claims that the agent should have resisted, or should have tried harder to resist, the influence of the subverting cause. I'll make two replies to this claim.

---

[31] In this third case, is doesn't matter whether Martha is aware of the relevant situationist research.

[32] In the counterfactual in which she knew all the causes of the action, she would learn nothing that would affect her F-acceptance. She already knows that the presence of the bystander is a cause.

[33] In the actual study, 93% of subjects offered no help to the researcher (see section 3.2.2).

[34] In the first explanation, the agent had not read about situationist research and she would not F-accept her action on learning of its causes.

[35] In section 5.1.2 I'll discuss a line of argument which is very similar to this objection, using another of the five key cases – that of Samantha who walks past a slumped man on her way to give a presentation.

[36] The full objection runs like this. If the agent is directly blameworthy, then she is DMR. If she also does not meet the CFA condition, then my conclusion that the CFA condition is necessary for DMR is refuted.

My first reply is that the notion of resisting an influence, though it features prominently in everyday action-explanation, is somewhat murky and difficult to elucidate. The assumption implicit in the objection is that resisting an influence is a process which is independent of the operation of the influence. But there are good grounds to doubt that. If resisting an influence is analogous to resisting a desire, then the very process of resisting may itself be influenced by a subverting cause. This is one of the lessons to be learned from research pioneered by Roy Baumeister. An agent's ability to resist a desire can be influenced by earlier incidents which have depleted the inner resource used for both "self-regulation" and logical reasoning.[37] For example if Gina had earlier completed a complex puzzle, she might now be unable to resist a desire to eat a cookie.[38] If she does not know of that effect, and would not F-accept the desire if she did, then the earlier puzzle completion is a subverting cause of her now eating a cookie – and also of her *failing to resist* her desire to do so.

Furthermore, resisting an influence is not something which can happen independently of the agent's desires. An influence can be resisted only if the agent desires sufficiently strongly to resist it. But a desire to resist something can itself be influenced by a subverting cause. Suppose that Martha desires to help the fallen researcher, but also desires to stay seated completing her questionnaire. Let's assume, for the sake of argument, that she can somehow try to resist the subverting influence of the person sitting next to her, despite not knowing of its influence. The success of her attempt to resist the influence will depend on the strength of her desire to resist it, relative to her desire to perform the action. But it's very likely that the strengths of both of those desires are influenced by the subverting cause. Therefore it is naïve to appeal to an effort of resistance in the face of a subverting cause, and unreasonable to blame someone for failing to succeed with that effort.

I'll now set aside that first reply to make a second, independent, reply. I think it's difficult to assess the claim that an agent should have tried harder to resist a subverting cause. We know so little about subverting causes, and their effects are so far removed from popular understanding, that we have no established and reliable intuitions about whether agents are blameworthy for not resisting them. For guidance I think we should compare them with other influences which agents are sometimes blameworthy for failing to resist, and about which we do have well established intuitions.

Suppose Paula is a police officer who is building a case against a man accused of assault. Paula is approached by the accused and threatened: Paula's beloved sports car will be smashed up unless she drops the case. She does drop the case. The objector will argue here that Paula is directly morally responsible for her action despite the coercive influence of the threat, because she should have done otherwise and is blameworthy. I think that is very plausible, and I don't dispute the argument in this case. Nevertheless, I think that actions with subverting causes are not analogous with Paula's action. I'll distinguish three types of influences over actions. Paula faces what I'll call a type-1 influence: *a known influence* to do the wrong thing. In type-1 cases, there is a clear moral onus on the agent to resist the influence, and a clear means by which she should resist it – in this case, by steeling herself to accept the possible consequence that her car may be smashed up.

Type-2 influences are more subtle. The agent *knows or suspects* that she is being influenced, but *does not understand how* the influence operates. For example, Rhona is discussing a possible repair to her roof with a tradesman. She knows that tradesmen sometimes employ influencing techniques, so she is on her guard against making a hasty decision which she might later regret. In fact the tradesman's techniques lead her to paying for a very expensive and unnecessary repair to her roof, and subsequently Rhona has to stop her donations to charity in order to pay for it. She has been influenced, but the objector

[37] See section 3.6.2.
[38] See section 3.2.5.

nevertheless will hold Rhona DMR for agreeing to the repair. According to the objector, there is a moral onus on Rhona to resist the influence, and some ready means by which she should be able to do so.

I am not convinced that there is always a means open to agents to resist type-2 influences. Furthermore, I think it's likely that the responsibility Rhona bears here is *traced*, not direct; in order to resist the influence she must take some earlier action, such as reading about how to resist persuasive tradesmen, or "getting herself into an extremely sceptical frame of mind". But I'll set aside those replies for the sake of argument, in order to meet the objector's case in its strongest possible form. Even if we allow that Rhona bears DMR for her action with a type-2 influence, I think we must deny that agents can be DMR when the influence is of type-3.

When an action has a type-3 influence, the agent *does not know or suspect that there is any such influence*. For example, Suzanne might be influenced to buy a certain brand of drink by a series of images in a TV advert deliberately intended to be processed subliminally.[39] As a result she does not buy the alternative drink which is ethically sourced. She is unaware of the images, and unaware that there might be any subliminal influence on her action. The objector will claim that Suzanne is directly blameworthy for her action: she should have resisted the influence to perform the morally worse action. This claim is mistaken: it is implausible to say that Suzanne *should* have resisted this influence. There can be no moral onus on Suzanne to resist an influence of which she is not aware and which she does not suspect.[40]

I think that type-3 influence is the appropriate analogy for *unknown and unsuspected* subverting causes. For example, when Martha completes her questionnaire despite the researcher's cries from the next room, she does not know or suspect that the presence of the bystander is a cause of her action. Similarly Samantha does not know or suspect that being told she was late was a cause of her walking past the slumped man. Nor does Ingrid suspect that the content of the word puzzle was a cause of her interrupting a conversation. And nor does Gina suspect that resisting eating cookies was a cause of her giving up on a problem. In all of these cases, it is false to say that the agent should have resisted the influence of the subverting cause.

Of course, not all subverting causes are unknown and unsuspected. When an arachnophobic agent runs out of the room, the presence of a large spider is a cause which is well known – and its subverting effect is also well known to her.[41] The objector may claim that an agent should resist *known* subverting causes, when there is a substantial moral imperative. My second reply does not cover known subverting causes; for those causes I point only to my first reply above.

4.3.4. The agent should have paused for further reflection

In this form of the objection, the claim is that an agent should pause for reflection in the presence of a potentially subverting cause. For example, when she hears the researcher's cries, rather than continuing to fill out her questionnaire, perhaps Martha should pause to reflect on what is the best course of action. Because she does not do so, the objector claims, she is directly blameworthy for the action and so DMR for it.[42]

---

[39] In a study by Karremans, Stroebe and Claus (2006), subliminal priming of the brand name of a drink influenced subjects to choose that brand if they were already thirsty.

[40] What if there were an earlier onus on her *to suspect* a possible influence in the present? As well as being implausible in Suzanne's case, this suggestion introduces the tracing principle. It's not relevant here, because I am concerned with *direct* moral responsibility.

[41] This cause is subverting because she *does not* F-accept her effective desire; knowledge of all the causes of her action would not change that.

[42] Since Martha does not meet the CFA condition in her action, this would refute my conclusion that the CFA condition is necessary for DMR.

My first response is that it's not obvious that failure to pause and reflect before an action is good grounds for holding an agent blameworthy for that action. But I won't rely on this line of response, and I'll set it aside.

My second response is that there may be a subverting cause of an agent's *not* pausing and reflecting about a certain action. An example might be that of Ingrid, who interrupts a conversation after completing a rudeness-themed word puzzle.[43] The words in the puzzle may be a subverting cause of Ingrid's failing to pause and reflect before interrupting.[44] Should we hold Ingrid blameworthy for *failing to pause*, on the grounds that she should have paused to reflect *about pausing to reflect* before acting? A vicious regress threatens.

My third response is to point out that pausing itself may be ethically wrong, in some situations. For example, if Penny is running late for a presentation she has promised to give, it may be ethically wrong for her to pause for reflection several times on her way. If she is travelling through a city on foot at rush hour, for example, there may be dozens of situations in which some subverting cause might be a cause of her failing to notice a moral demand. But if she pauses for reflection dozens of times on her journey, to assess whether there *might* be a moral demand that she should act on, she may well be late for the presentation. It's not clear that pausing for reflection is the morally best thing for her to do. I think it would be false to say Penny is blameworthy for any particular action of continuing on to give her presentation, on the grounds that she does not pause to reflect. So it seems that there cannot be a general principle that people should pause to reflect before acting, even in situations where there might be unnoticed moral demands.

This links to my fourth response: the injunction to pause and reflect at any occasion in which there is a possible subverting cause is simply impractical, if my claim that subverting causes are commonplace is correct. It's not plausible to think that we are repeatedly blameworthy for failing to pause on every such occasion.

This point in turn links to my fifth response. In order to be blameworthy for an action in virtue of failing to pause and reflect, the agent must have some reason to think that she *should* pause and reflect. She must consider that her initial assessment of the situation may be flawed, and that further reflection might be warranted. For example, when Martha is filling out her questionnaire, she assesses the researcher's cries as not indicating that help is needed. For what reason would she now pause to reflect before continuing to complete the forms? She would only do so if she suspected that her initial assessment may be distorted. But she would only suspect that if she also suspected that some element in the situation might be a cause of a distorted assessment (in this case, the presence of a bystander). And of course, she does not suspect that, since she has not read the relevant research literature. So Martha has no reason to think that she should pause before acting. Hence she is not blameworthy in virtue of not pausing.

It may be objected, to that line of argument, that Martha has reason to pause even though she does not know about situationist research. It makes sense, the objector may say, that something like "peer pressure" or social influence might affect her action, and hence she is indeed blameworthy for not considering that possibility and pausing to reflect. To this objection I reply that, though we as everyday agents may have some knowledge that social influence can affect decision-making, we do not expect it to affect our *assessments* of moral demands. If we are aware of social influence as a phenomenon at all, we tend to think that it operates by making us consciously too embarrassed or inhibited to perform certain actions.[45]

---

[43] See section 3.2.4.

[44] In the absence of the cause, she would have paused to reflect. But if she knew of the cause, she would not F-accept her failure to pause.

[45] In this respect the objection mirrors our limited everyday understanding of these phenomena. We are inclined to object that Martha should have paused to reflect because she must have felt embarrassed or inhibited. But it need not be the case that she felt those emotions. (See also section 3.8.5.)

Furthermore, even if we have *some* understanding that social influence might affect even our assessments of moral demands, we certainly do not expect it to have as profound an effect as it does: we do not imagine that sitting next to a passive bystander would lead Martha to not to react to the researcher's cries, when 70% of people sitting *alone* do offer help.[46] I don't think it's plausible to blame Martha for not pausing to reflect, in this situation.

Furthermore, Latané and Darley – who ran a series of studies into factors inhibiting bystander intervention – discovered that pausing to reflect in situations like Martha's actually made it *less* likely that the agent would intervene to help.

> Subjects in our experiments responded early or not at all. Over 90 percent of all subjects who responded responded within the first half of the relatively short time available to them … By their initial inaction, [many other] subjects inadvertently committed themselves to continued inaction.[47]

Meanwhile, suppose that someone in Martha's position (I'll call him Malcolm) also assessed the researcher as not in need of help when she cried out. Unlike Martha, Malcolm *did* wonder whether his assessment was influenced by the bystander's presence. Could Malcolm be blameworthy for his action in virtue of failing to pause and reflect before performing it? Given his set of circumstances, I think that is plausible. But Martha's case need not be like Malcolm's.

A final response is that, in many actions which have subverting causes, there is even less reason for the agent to suspect that a pause for reflection might be beneficial. For example, when rating the qualities of job applicants, we do not expect that our ratings will be affected by whether we will later meet the applicants.[48] There is no apparent reason for someone making the ratings to pause and reflect, and no grounds to hold her blameworthy in virtue of not pausing. It's therefore even less plausible to blame Judy for not pausing to reflect about her rating, since the cause is even more unexpected than the cause in Martha's case.[49]

4.3.5. <u>There is a flaw in the agent's character or dispositions</u>

The final objection in this series is the claim that the agent is directly blameworthy for her action (and hence DMR for it) because some aspect of her character or dispositional nature is weak or flawed. For example, the objector might say of Martha that she is more readily influenced than most by the behaviour of other people. This flaw, he claims, is what explains why Martha failed to help the market researcher, while 7% of people in the same situation did help.

Before replying to this objection, I'll point out again that it would not be made by an identificationist like Frankfurt. For Frankfurt, it is necessary for responsibility that the action and effective desire were truly the agent's own. If they were not then the agent is not morally responsible, no matter what can be said about her character and dispositions.

Setting that point aside, I'll make two points in reply to the objection. My first is that there is no evidence that the agents who failed to help in Martha's position do indeed lack or possess any particular character or dispositional trait. As I discussed in section 3.7.3, studies of this kind often test for such correlations and fail to find them. It may be that there is no trait which correlates significantly with performance of the action, counterintuitive though this seems. By a "trait" here I mean some feature which is exhibited across many situations. But

[46] See section 3.2.2.
[47] Latané and Darley 1970:122. See also section 3.8.5.
[48] See sections 3.2.3 and 3.8.1.
[49] Similar points can be made for Samantha, Ruby and Gina – the other agents in the five key cases.

situationism "denies that people typically have highly general personality traits that effect behavior manifesting a high degree of cross-situational consistency".[50]

My second point is about the structure of the objector's claim here. How can an agent be directly blameworthy for an action *in virtue of the fact that* her character or dispositional nature is flawed? It seems to me that two conditions must be fulfilled. The agent must know or suspect that the flaw is present; and it must be reasonably foreseeable to her that performing the bad action would result from that flaw.[51] There is no good reason to think that the first condition is fulfilled in Martha's case. This would require that there is some broad character trait (such as being more readily influenced than most by the behaviour of others) which correlates significantly with performing the morally bad action in this particular situation – for which there is both no evidence and also some grounds for doubt. It would also require that Martha is aware that she lacks that trait, which might well not be the case.

More importantly, the second condition is almost certainly not fulfilled in Martha's case: it is not reasonably foreseeable to her that being more readily influenced than most by the behaviour of others would lead her to fail to respond to a researcher's cries in the very next room. If the contrary were the case, we would not be as shocked as we are to learn that only 7% of people in her situation did respond.

Suppose that the psychologists had *described* the situation to a control group of subjects, who were not actually placed in that situation. I think that the great majority – far more than 7% – would predict that they would offer help to the researcher. It is surely not reasonably foreseeable to these would-be subjects that they would perform the bad action as a result of possessing a certain dispositional trait, if they are untrained in social psychology. After all, in such experiments even expert social psychologists often fail to uncover any dispositional traits which correlate with performance of actions. And yet the trait (or traits) would surely have to be extremely common, to explain the result that 93% of people will fail to help.

I conclude that it's unreasonable to hold Martha directly blameworthy and DMR for her action in virtue of possessing a certain trait, because she cannot reasonably foresee that possession of the trait would lead her to perform that action (if indeed there *is* a trait which correlates with performing that action). Therefore the objection does not refute my proposal that the CFA condition is necessary for DMR.

I'll now turn away from the series of objections which seek to establish that the agent in one of my five example cases is directly *blameworthy*, and thereby DMR for her action. The final objection aims to establish that the agent is DMR by different means.

4.3.6. The agent would feel responsible despite subverting causes

Sometimes, even after she has learned of the subverting cause of her action, an agent may *feel herself to be morally responsible* for that action. For instance, Martha might feel morally responsible for continuing to fill out her questionnaire when the researcher cried out, despite learning that the passive behaviour of the person sitting next to her was a cause of her action (see section 3.2.2).

Suppose that Martha is told of her action's causes, and feels ashamed and angry at herself. These are emotions suggestive of someone who feels responsible for her actions. Even if she were told of various situationist studies which reveal the surprising influence of passive bystanders, she might remain ashamed or angry. This indicates that Martha would not

[50] Doris 2002:38-39.

[51] Arguably, there is a third necessary condition: the agent must have been able at some earlier time to change her nature to remove the flaw. But I'll set this condition aside since it is more controversial than the other two.

F-accept her action at the time of acting, if she knew its causes.[52] Yet Martha feels morally responsible for performing the action. This is evidence, the objector claims, that she is directly morally responsible for her action – despite the fact that she would not F-accept it.

I'll make three points in response to this objection. The first point is that *feeling* morally responsible is not the same thing as *being* responsible. I'm working on the assumption that an agent's moral responsibility is a matter of her *being an appropriate candidate for* reactive attitudes in relation to the action, independently of any facts about whether the action in fact stimulates any reactive attitudes.[53] One of my aims is to show that we are often mistaken when we feel ourselves and others to be morally responsible for actions.[54] Thus I regard Martha's feeling morally responsible for her action as *prima facie* evidence of her actual moral responsibility, but not *decisive* evidence.

Nevertheless it's surely true that most people in Martha's position would feel similar reactive attitudes toward their own action. And in fact most onlookers would hold her morally responsible for her action, if they were not familiar with situationist research. This adds up to quite a strong prima facie case for Martha's moral responsibility. My second point is that Martha's own feeling that she is morally responsible could be overturned. It may seem paradoxical that a philosopher or a psychologist looking at the case is less likely to hold the agent morally responsible, in that situation, than the agent herself. But the paradox dissolves, I think, when we pinpoint the difference between the people making these judgements.

If the difference were that the philosopher (or psychologist) had very different fundamental intuitions about moral responsibility from the agent's, then that would be problematic, and we might suspect the philosopher of having lost touch with something important about the concept. But that is not the case here.

Instead, the difference is simply a matter of information, as we can see if we consider a series of steps in which the agent receives more and more information. Martha does not know all the causes of her action, so it is natural that she feels morally responsible for it. If she were told of the subverting cause, she might feel slightly inclined toward excusing herself, but not convincingly so. If she were told that only 7% of subjects in her situation helped the researcher, she would probably feel more inclined to excuse herself. If she then studied situationist psychology, and learned that the same pattern has been found in many similar experiments, she would realise that agents can be very powerfully influenced by situational factors, and be more inclined again to excuse herself. If she then considered the philosophical implications, as discussed here and by Doris and Nahmias (see chapter 5), it's more likely still that she would no longer consider herself to be morally responsible for her action.[55] None of this extra information would change Martha's intuitions *about the concept of moral responsibility*. Rather, it would change her view about *which actions* fall within the scope of the concept.

My third and more minor point is that Martha's feeling that she is morally responsible for her action need not indicate that she feels *directly* morally responsible for it. She might instead feel that she bears traced moral responsibility for it. She might continue to feel this, even after learning of the subverting cause of her action. For in our feelings about whether someone bears moral responsibility we do not usually distinguish between direct and traced moral responsibility. The feeling is the same in either case. Martha would be TMR if the resultant occurrence of her action was reasonably foreseeable by her at the time of some

[52] If Martha *would* F-accept her action, then she would be directly morally responsible for it on my account – and so the objection would not be applicable.

[53] This is Fischer and Ravizza's definition of moral responsibility, which I adopted as a working assumption in section 1.1.3.

[54] This will be the subject of chapter 7.

[55] Of course, I can't claim that everyone would come to this conclusion, but I think it's likely that many would do so. Furthermore, this conclusion would not seem completely implausible, even to people who did not in the end agree with it.

earlier action or omission. In her particular case, it's very unlikely that there is any such earlier time, since she does not have specialist knowledge of psychology. But judgements about traced moral responsibility are sometimes complex and difficult to make. We are used to feeling morally responsible even for actions with which we do not identify, in part because we accept the tracing principle and think that it often operates in everyday actions. We are often reluctant to excuse agents from responsibility when there is uncertainty, especially when their action was bad, or the consequences of their action were bad.[56] For Martha, the possibility that she might bear TMR may explain her feeling morally responsible for her action. Thus Martha's feeling that she is morally responsible does not necessarily indicate that she feels *directly* morally responsible for it.

I conclude that the objection fails. That an agent would *feel herself* to be morally responsible for an action with subverting causes, *even if she were told* of the role of the subverting cause, does not show that she is DMR for it. In fact, this consideration is outweighed by the arguments which point to her lacking DMR for it.

4.4. CONCLUSIONS

I've considered three kinds of objection to my arguments against Frankfurt.

In section 4.1 I discussed situationist causes which do not subvert moral responsibility on my account, because agents would continue to F-accept their actions on learning of those causes. For example, finding a dime in a phone booth may be a non-subverting cause of a helping action. I considered a series of objections which suggest stronger necessary conditions of DMR than my CFA condition. All of these conditions imply more sceptical conclusions than mine about the prevalence of moral responsibility for actions. Although these conclusions are not threatening to my position, I argued that all of them can plausibly be rejected by compatibilists.

In section 4.2 I considered cases of regrettable actions for which it is obvious that the agent is DMR, and yet it appears that my CFA condition is not met. These would be counterexamples to my claim that the CFA condition is necessary for DMR. However, I showed that the CFA condition is indeed met in these cases, and so the objection dissolves.

In section 4.3 I discussed a series of objections with a common structure: an agent who does not meet the CFA condition might seem to be DMR for her action, in various different ways. If one of these objections succeeded, it would show that the CFA condition is not necessary for DMR. The objectors' points are most persuasively applied to Martha's action in the 'market research' case, so I focused on that case. I argued that none of the objections establishes that Martha is DMR for her action. I think the same points would be equally effective in the other five cases I rely on in my main argument.[57] Meanwhile I need only one case in which an agent is not DMR for her action, while she does F-accept her effective desire, in order to establish my main argument against Frankfurt.[58]

In this chapter I've defended my two arguments against Frankfurt's position (my main argument, and my positive proposal). Very similar objections could be made to the arguments I'll make in the next two chapters, against other theories of moral responsibility. My replies, too, would be very similar to those I've given here.

[56] Empirical research by philosophers seems to show that people blame agents for the bad consequences of their actions more readily than they praise agents for good consequences – see for example Knobe 2006.

[57] In fact, I think my points would be *even more* plausible when applied to the other four cases, though I have not argued for that.

[58] The conclusion of my main argument is that F-acceptance is not sufficient for DMR. I discussed this point in more detail at the start of section 4.3 above.

I'll turn next to discuss two philosophers who have specifically addressed the threat to moral responsibility posed by situationist social psychology: John Doris and Eddy Nahmias.

5. DORIS AND NAHMIAS

In chapter 3 I set out my main argument against Frankfurt's theory of moral responsibility. I concluded that Frankfurt's condition of identification is not sufficient for direct moral responsibility.[1] I argued that when an action has subverting causes, the effective desire which moves the agent to act is not truly her own.[2] On an axiom of identification which Frankfurt himself holds, an agent cannot be directly morally responsible for an action if the effective desire is not truly her own.[3] I also made a positive proposal of an alternative condition which I believe is necessary for direct moral responsibility.

John Doris and Eddy Nahmias are two philosophers who have engaged in detail with situationist research and its implications of for moral responsibility. They both propose identificationist theories, which I will discuss and criticise in this chapter.

My five key cases of subverting causes included three situationist factors –Samantha's being told that she is running late for a presentation (section 3.2.1), a bystander sitting next to Martha as she filled out a questionnaire (3.2.2), and Judy's being told that she will meet the candidate whose qualities she is assessing (section 3.2.3). Doris and Nahmias do not discuss the other kinds of subverting cause which I included in section 3.2.

I'll begin by addressing Doris's account.

5.1. JOHN DORIS

Doris's book *Lack of Character* is focused primarily on the relationship between character and ethical behaviour, but it includes one chapter about responsibility.[4] In this section I will summarise Doris's arguments for his proposed condition of responsibility.[5] In section 5.1.1 I'll criticise Doris's arguments and explain why I think my condition is preferable. In section 5.1.2 I'll summarise and then criticise Doris's conclusions about responsibility. Finally, in section 5.1.3, I'll summarise my conclusions about Doris's theory.

Doris aims to develop an identificationist theory which can resist the threats to our attributions of responsibility posed by situationist research. He does not examine or criticise any previous identificationist account.[6] Instead, he begins by outlining a very *generic* condition for what I'll call "G-identification":

> To identify with one's determinative motive is to embrace it or regard it as "fully one's own".[7]

Doris speaks of agents being identified with "determinative motives", whereas Frankfurt-style identification is with "effective desires".[8] Doris's "determinative motives"

[1] Direct moral responsibility (DMR) is anchored in facts about the action and the agent at the time of acting. In contrast, traced moral responsibility (TMR) can be "traced back" to facts about an earlier action (or omission) and about the agent at that earlier time. See section 1.2.

[2] An action has subverting causes when it is true that: if the agent knew all the causes of her action, she would not F-accept the action. (This is a simplified definition; for the full version see section 3.1).

[3] This axiom is statement (4) in the set which define Frankfurt's position in section 2.2.

[4] The following caveat applies. Chapter 7 of *Lack of Character* covers a lot of ground, and Doris admits engagingly that he has not filled out all of the details for some of the positions he considers and adopts. It's therefore possible that I've misunderstood his views on some points.

[5] Doris speaks of "responsibility", but by that he seems to mean what most other philosophers mean by "moral responsibility". At any rate, I don't think it makes a significant difference to my arguments if the two concepts do not have exactly the same extensions.

[6] Doris notes "I favor the "identificationist" approach to moral responsibility associated with Frankfurt"; but he does not describe Frankfurt's account in detail (2002:140).

[7] Doris 2002:140.

[8] Sometimes Doris speaks only of identification with a "motive" for action; I assume that he uses this term as shorthand for "*determinative* motive".

include desires which would be classed as "endorsing higher-order desires" on Frankfurt's account.[9] For example, on Doris's account, an agent can be identified with a determinative motive to please her father. The same agent on Frankfurt's account could be identified with an effective first-order desire to perform a specific action, but the desire to please her father would be among the higher-order desires which endorse that first-order desire. Despite this difference, I think the two theories are readily comparable. To assess moral responsibility for a particular action, we are interested in whether the agent identifies either with the effective desire (Frankfurt) or with a determinative motive which endorses the effective desire (Doris). For consistency with previous arguments I will sometimes express my points in terms of "effective desires" rather than "motives". In doing so I don't think I risk any misunderstanding of my arguments.

Ignoring situationist experiments to begin with, Doris argues that G-identification is neither necessary nor sufficient for moral responsibility. That G-identification is not necessary for moral responsibility is shown, according to Doris, by the example of an agent who is ashamed of her akratic action when betraying her spouse.[10] This is an action in which the agent is not G-identified; and yet the lack of G-identification does not absolve her from moral responsibility.[11]

Meanwhile, to see that G-identification is not sufficient for moral responsibility, we need only consider cases where the agent suffers a deprived upbringing or a mental illness, or has been brainwashed.[12] Necessary for moral responsibility, Doris believes, is "normative competence", which is

> a complex capacity enabling the possessor to appreciate normative considerations, ascertain knowledge relevant to particular normative judgments, and engage in effective deliberation.[13]

Via "effective deliberation" the agent can make decisions which conform with her evaluative commitments.

Doris aims to specify a form of identification which is necessary and sufficient for responsibility. Doris considers two issues which threaten responsibility when actions have situationist causes. The first issue he considers is that agents whose actions have situationist causes *do not exercise their capacity for effective deliberation*. The second issue is that agents whose actions have situationist causes *have not consciously recognised their own motives for acting*.[14] Doris argues that neither of these issues prevent an agent's moral responsibility for her action. He employs the same example to deal with both issues. When we act out of habit and without reflection, we sometimes do not exercise our capacity for effective deliberation, and yet we are often morally responsible. And when we act out of habit and without reflection, we sometimes have not consciously recognised our own motives, and yet we are morally responsible. Thus neither the exercise of effective deliberation nor conscious recognition of motives are necessary for moral responsibility. Hence Doris also concludes that

[9] Motives may also include beliefs about which first-order desires will best achieve the higher-order desire.
[10] Doris 2002:140.
[11] For comparison, it's not clear to me whether Frankfurt (in his final account) regards identification as necessary for moral responsibility, and I did not assume that he does. See section 2.2.
[12] Doris 2002:216n37.
[13] Doris 2002:136. Doris credits Susan Wolf for inspiring this part of his account. I discussed Wolf's criticism of Frankfurt on this point in section 2.3.1.
[14] Doris uses the expression "consciously identified motives" (2002:141) in place of "consciously *recognised* motives", but I think that phrasing is apt to confuse in the context of a discussion about identifying *with* motives.

neither of these is necessary for identification, since identification is necessary for moral responsibility.[15]

Instead Doris briefly contemplates the following identification condition involving counterfactual scrutiny:

> Identification may be said to obtain if a person would have [G-] identified with the determinative motive of her [action] at the time of performance had she subjected it to reflective scrutiny.[16]

Reflective scrutiny here involves both the exercise of effective deliberation and conscious recognition of one's own motives.[17] But Doris rejects this condition with the following argument.

Suppose an agent (I'll call her Fatima) has an unconscious desire to please her father, which influences her action. If she reflected on her action, she would not recognise that this unconscious desire played any role in it. If she were made aware of this desire, at the time of acting, she would recoil from it, saying "Well, if *that's* why I'm doing it, I'm damn sure going to do something else".[18] According to Doris, this lack of counterfactual identification need not excuse Fatima from moral responsibility. If Fatima's unconscious desire for paternal approval has influenced many of her actions, including several which are harmful to her colleagues for example, those colleagues may indeed hold her morally responsible for her latest action. Therefore the counterfactual scrutiny condition is not necessary for moral responsibility (and so it is not the correct condition of identification).

Doris's response to this problem employs what he calls "narrative integration". A motive admits of narrative integration if it can be "integrated into a narrative that manifests identification".[19] A narrative can reveal identification even when the agent (the subject of the narrative) disavows that motive; it can reveal that the motive does in fact express the agent's "operative priorities" or "evaluative orientation". The agent's judgement about whether she is identified with a motive is not definitive:

> Narrative development is a way of subjecting claims of identification (or its lack) to the sort of interpersonal scrutiny required for responsibility assessment; this may sound too impressionistic for good philosophy, but I think it is a fair characterization of something central to our practice, and it is far from clear that there is a better way to proceed.[20]

I summarise Doris's argument in a condition of identification:

> *NI* An agent identifies with a determinative motive to perform an action if and only if that motive admits of narrative integration.

On condition NI, Fatima may well be identified with her unconscious desire to please her father – if a narrative reveals that it does express her operative priorities or evaluative orientation.

---

[15] For Doris, moral responsibility requires identification. If it were the case that identification requires effective deliberation, for example, then we could conclude that moral responsibility requires effective deliberation. But moral responsibility does not require effective deliberation. Therefore identification does not require effective deliberation.

[16] This phrase is a quotation from Doris 2002:141.

[17] At least, this is my interpretation of Doris's remarks on the subject (2002:141).

[18] Doris 2002:141. Doris does not give an example of the kind of action he has in mind; I suggest one in the discussion below.

[19] Doris 2002:142.

[20] Doris 2002:142.

I take it that Doris regards NI as necessary and sufficient for moral responsibility.[21] What verdict does the NI condition give about actions with situationist causes? Consider for example the case of Samantha, who walks past a slumped man after being told that she is running late to give a presentation.[22] Doris acknowledges that there are some important differences between unconscious motives and "the kind of motives adduced by situationism". Fatima's unconscious desire to please her father is a very *personal* motive; whereas Samantha's desire to be on time for her presentation seems rather *generic*. Fatima's desire is representative of a long history of similar actions, and fits into a temporally extended narrative. It's not so easy to fit Samantha's desire into a narrative; actions like hers "look like psychological tics or glitches".[23]

Doris now appeals to the notion of "policies" – which "reflect ongoing commitments that order and structure behavior".[24] Take the example of a "hard-charging stockbroker" (I'll call him Steve) who walks past someone in distress without helping, where a cause of Steve's action is that he has been told that he is late for a presentation.[25] Steve might meet Doris's identification condition NI, for

> even where a person fails to identify with the callousness that resulted from haste, he might yet embrace having the sort of packed schedule that induces haste. The driven stockbroker might reject his many hasty omissions, but wholeheartedly endorse the way of life that leads to them. The hard-charging broker embraces a life-plan that eventuates in the unfortunate omissions; the identification requisite for responsibility obtains, although it is not directly proximate to the omission.[26]

Doris discusses an interesting and difficult problem case in which an agent (I'll call him Alf) performs an akratic act of sexual infidelity under intense situational pressures brought to bear by "a committed and skilled seducer".[27] The sexual infidelity represented no policy of Alf's; in fact it happened *despite* his policy to the contrary. However, as Doris notes, "the proximate motive is not the only motive of ethical interest".[28] If Alf's akratic infidelity follows his wholehearted flirting during a candlelit dinner, then in this "there is a legitimate target for the reactive attitudes, though it comes rather earlier in the process than we might have expected".[29] Doris goes on to suggest that Alf is responsible for the flirtation, but not for the infidelity – despite the likelihood that his wife will feel much angrier about the latter than the former. Similarly he suggests that someone may be responsible for drinking too much alcohol, but not for his subsequent behaviour.[30] Doris recommends taking this view as an improvement on our "familiar reactive habits".

---

[21] Doris holds that effective deliberation is necessary for normative competence, normative competence is necessary for identification, and identification is necessary for moral responsibility (2002:136 and 140). Like me, Nelkin interprets Doris as holding that the condition is sufficient for moral responsibility: "One is responsible for an action [according to Doris] if one acts on motives that admit of narrative integration" (2005:197).

[22] See section 3.2.1. Doris doesn't discuss individual actions, but speaks only in general terms about "situationally induced behaviors".

[23] Doris 2002:143.

[24] Doris 2002:144. Doris also uses the word "plan" as a synonym for "policy". I find this confusing, so I've used only the term "policy".

[25] Doris does not say explicitly that the example of the hard-charging broker is supposed to shed light on actions with *situationist* causes, but that seems to be implied by the position of these remarks in his argument.

[26] Doris 2002:144. The sense in which the person "fails to identify" in the first sentence of this quote is presumably the generic sense I labelled "G-identification".

[27] Doris 2002:144. (The example is embellished at various points during the chapter.)

[28] Doris 2002:144.

[29] Doris 2002:144.

[30] Doris 2002:145.

Doris concludes that, far from undermining responsibility, situationist findings can *increase* responsibility, by enabling more effective deliberation.[31] Those who know about situational stimuli are *more* responsible than they were before gaining that knowledge. He also holds that agents are often morally responsible even when their actions result from situationist effects of which they are not aware. I'll discuss those conclusions in more detail in section 5.1.2, but first I will criticise Doris's responsibility condition, and explain why I think my proposal is preferable.

5.1.1. Criticisms of Doris's account

I'll begin by amending my proposed necessary condition for DMR, to incorporate the generic notion of identification with which Doris compares his account. I labelled that notion "G-identification":

> To identify with one's determinative motive is to embrace it or regard it as "fully one's own".[32]

The condition which I argued for in section 3.5 was this:

> *CFA* The agent *would F-accept her effective desire*, if she knew all the proximate and relevant causes of the action.[33]

An equivalent condition, incorporating generic G-identification in place of Frankfurt's F-acceptance, would be this:

> *CFG* The agent *would embrace her determinative motive or regard it as fully her own*, if she knew all the proximate and relevant causes of the action.

I'll argue, against Doris, that this CFG condition is necessary for direct moral responsibility.

There are several important points on which I agree with Doris. First, I agree that G-identification is not sufficient for moral responsibility: normative competence, including the agent's ability to deliberate effectively, is also necessary. Second, I agree that the agent's conscious recognition of her motives is not necessary for moral responsibility (whether DMR or TMR). The key question, in my view, is whether the agent would embrace her determinative motive, or regard it as fully her own, *if she knew all the causes* of the action.[34] Third, I agree that the *exercise* of effective deliberation is not necessary for moral responsibility, nor for direct moral responsibility. An agent's motive can be fully her own when it is unconscious, and also when she acts out of habit without reflecting on the motive. On my account, a motive can be fully the agent's own when effective deliberation does not occur because the action has a situationist cause, provided that she would embrace it if she knew of the situationist cause.[35]

I also agree with Doris that we can use knowledge about situational effects (and indeed other types of subverting cause which he does not discuss) to make better choices (both

[31] Doris 2002:129 and 153.
[32] Doris 2002:140.
[33] This is a simplified version of the condition; see section 3.5 for the fully precise version.
[34] The weaknesses of the phrases "embrace" and "regard as fully her own" are evident, but I've used them because they are part of Doris's rough and generic characterisation of identification. It strains normal usage to say that an agent can "embrace" a motive that she does not consciously recognise, or "regard it as fully her own". A more nuanced notion is needed instead – such as Frankfurt's F-acceptance.
[35] In this kind of case the situationist cause is not subverting. I discussed one such case in more detail in section 4.1.

ethically and prudentially better).[36] However, I think Doris is mistaken in his assessment of agents' responsibility when their actions have situationist causes which are subverting. Standing back from his theory, one might view it as moulded with the end in mind: Doris wants his theory to give the verdict that agents are usually morally responsible for actions with situationist causes.[37] Although it achieves that end, the theory fails to deal effectively with actions whose causes are already familiar to us.

In the important case of Alf's akratic act of sexual infidelity, Doris claims that Alf bears moral responsibility only for his earlier flirting over dinner.[38] I think it's much more plausible to say that Alf bears moral responsibility *for his act of infidelity* (as well as for his flirting). Doris acknowledges that adopting his proposal would require a change to our reactive practices.[39] I see no need for this change, given that an obvious mechanism is available by which we can hold Alf responsible for the act of infidelity – namely tracing. Alf bears traced moral responsibility for the act of infidelity because the earlier flirtation is fully his own, and because the subsequent infidelity is reasonably foreseeable by Alf at the time of the flirtation. Doris makes a comparison between this case and that of a man who drinks too much alcohol and later behaves badly. Again it seems to me that we hold the drunken man responsible *for his actions while drunk*, and not merely for drinking too much. He bears TMR for his actions while drunk. Doris's proposal, that we excuse Alf and the drunk for responsibility for their lustful and drunken actions, goes against our reactive attitudes and intuitions. I think my proposal is much less radical: we should accept that Alf and the drunk are not *directly* morally responsible for the lustful and drunken actions, because these actions are not fully their own. But the agents are morally responsible for them nonetheless, via tracing.

An advantage of Doris's account of responsibility – compared with Frankfurt's – is that the relevance of the agent's history is recognised. The means by which it is recognised in Doris's account is "narrative integration". Doris acknowledges that using narrative integration as the basis for judgements of responsibility will be difficult in many cases. Often our judgements will involve "more art than science".[40] I don't regard this as a weakness of his account. Judgements of moral responsibility are often difficult to make.

A more serious problem is the threat of circularity in his account. It's far from clear that our judgements about whether an agent's desire to act admits of narrative integration are *independent* of our judgements about whether the agent is morally responsible for an action. According to Doris, in obsessive-compulsive disorders there may be desires which are very typical for the agent which nevertheless do not meet the narrative integration requirement.[41] Suppose that Oscar washes his hands repeatedly after eating and cannot bring himself to help with cleaning the dishes. We do not hold Oscar morally responsible for his obsessive hand-washing action because, according to Doris, we cannot readily integrate the desire for this action into a narrative. But to see a problem with this view, I think we need only think back to a time when obsessive-compulsive disorders were not understood. Without an understanding of the disorder, people might well have held Oscar morally responsible for his hand-washing action. The associated narrative might have described Oscar as overly and selfishly worried about his own cleanliness, or habitually lazy and unwilling to help with household chores.

[36] See Doris 2002:153.

[37] This impression is reinforced by Doris's insistence that exculpation for those who are unaware of situational effects is a result "to be resisted" (2002:153). I'll discuss this point further in section 5.1.2.

[38] At least, this is how I interpret Doris's remarks when he says that "there is a legitimate target for the reactive attitudes, though it comes rather earlier in the process than we might have expected" (2002:144). Doris makes no use of tracing, despite discussing several cases in which it seems natural to do so.

[39] Doris 2002:145.

[40] Doris 2002:145.

[41] Doris 2002:142. Doris does not support his point with any examples; the example of the hand-washing obsessive is mine.

With greater modern understanding of the condition, we realise that those narratives were flawed. But it cannot be that Oscar is excused from moral responsibility today because our narratives are accurate, whereas a medieval man in the same situation *was* morally responsible because his contemporaries' narratives were flawed. An agent's metaphysical status as morally responsible must not vary according to the narratives of the day.[42] (In contrast, my condition is not threatened by this circularity. In the counterfactual scenario in which the agent must embrace his effective desire, he knows the true causes of his action. If does not embrace it or regard it as fully his own, he is not DMR on my account.)

Perhaps Doris might respond to this objection by appealing again to the notion of policies.[43] He might claim that Oscar has no policy to wash his hands repeatedly after eating, and hence that the desire to do so is not suitable for narrative integration. But I don't think this move would solve the problem facing Doris. For Oscar might indeed have a deliberate policy to wash his hands repeatedly after eating, as a means of managing his disorder. Setting such a policy might enable him to wash his hands much less often during the rest of the day. So it may be that Oscar's having a policy in favour of his action makes no difference to whether he is identified with the desire (for Doris will surely still hold that Oscar is not identified with it here).

If Doris now responds that the relevant sense of a 'policy' covers only actions which are embraced by the agent as fully his own, then the notion of a policy itself appears not be independent of the notion of identification.

Furthermore, in times before our modern understanding, people might well have regarded Oscar's hand-washing actions as reflecting a policy (e.g. a policy of avoiding chores, rather than a policy of managing his condition). Those people would have been licensed by Doris's account to hold Oscar morally responsible – contra Doris's own verdict on this kind of case.

A different way to put the point against Doris is that the notion of "admitting of narrative integration" is vague. It's not clear why some obsessive-compulsive desires do not admit of narrative integration, if they are often effective and lead to action. Similarly if there are two kinds of policies, some of which reveal identification while others are merely compulsive, it's not clear what makes the relevant difference. That a certain verdict matches our practices of responsibility attribution does not provide it with properly independent support.

I think there is another problem with Doris's narrative integration condition, which he does not address. Suppose Karen is habitually extremely health-conscious and well disciplined, and has a policy not to eat unhealthy desserts in restaurants. In fact, she has never done so before. On this particular evening at the restaurant she has perused the dessert menu, and knows that it would be best, all things considered, to order a fruit salad. However, on this particular evening she has a strong desire to eat chocolate cake. She orders some chocolate cake. This action is akratic, but she does G-identify with her motive – she embraces it as fully her own. It seems that the verdict of Doris's account, though, must be that Karen is not identified with her motive, and so not morally responsible for her action. Her determinative motive is not readily integrated into a narrative, since she has never before ordered cake (nor any other unhealthy dessert) in such a situation. Furthermore, her action contradicts her explicit policy. But the verdict that Karen is not morally responsible for ordering the cake makes Doris's condition much less plausible than the generic condition of G-identification in this case.[44]

---

[42] On the view of moral responsibility which I adopted in section 1.1.3, a morally responsible agent is *an appropriate candidate for the reactive attitudes*. There is a fact of the matter about this, whose truth is not merely relative to the norms of a particular society.

[43] Doris does not mention policies in connection with this kind of case.

[44] On Frankfurt's account, too, Karen would be held DMR if she F-accepts the action as her own.

Doris might respond by claiming that Karen's action *is* readily integrated into a narrative: namely, the narrative I gave in the previous paragraph. But in order to assess whether the description in previous paragraph establishes Karen's identification with her motive, in the way that Doris needs, we must strip out the stated fact that *Karen does embrace it as her own*.[45] Without that fact, I cannot see any basis on which we might integrate Karen's strong desire to eat chocolate into a narrative. She has never before ordered an unhealthy dessert in a restaurant, and has a policy against doing so. And yet it may be that she does embrace the motive as her own. On that basis Karen is G-identified with her motive.[46]

Instead, perhaps Doris might respond that we must consider the *type* of action in this case. We might for example count Karen's action as of the type "indulging herself", and it might be that she has performed actions of this type before, though perhaps only rarely.[47] The problem with such a reply is that the threat of circularity lurks nearby. How are we to determine the types of actions which are relevant for narrative integration? How many times must an agent have performed an action of that type in order for her current motive to admit of narrative integration? Our answers to these questions cannot be guided by our intuitions about responsibility attribution, without circularity.

Furthermore, it's not clear from Doris's discussion what verdict we should give if an agent's narrative contradicts his policy. Suppose that Sidney, a smoker for the last 30 years, decides to stop smoking. I think it's reasonable to say that he now has a policy not to smoke, reflecting his ongoing commitment. Two days later, at a moment of extreme stress, he succumbs to his craving to smoke another cigarette. Should we say that Sidney is identified with his effective desire, on the grounds that it meshes well with his narrative? Or should we conclude that he is not identified, because the action contradicts his policy? At best there seems to be a lacuna in Doris's account here. At worst there may be an irresolvable tension.

Meanwhile, the narrative integration test does not always pick out desires with which the agent is identified, or which are fully her own. If an agent would disavow a motive if she subjected it to reflective scrutiny, I think the most plausible conclusion is that identification does *not* obtain.[48] I am not persuaded by Doris's key example of the agent with an unconscious desire (whom I called Fatima), which is supposed to show the contrary.

Fatima has an unconscious desire to please her father, which is a determinative motive in her action. Suppose that the action is indiscreetly revealing a secret about one of her colleagues, which in turn enhances Fatima's own prospects of promotion at her colleague's expense.[49] If she were made aware of the unconscious desire, at the time of acting, she would disavow it.[50] Doris believes that Fatima is identified with her desire if it can be integrated into a suitable narrative. I think this is an implausible view. I'll argue that, if Fatima would disavow the desire, then she is not identified with it. Yet, at the same time, that lack of

[45] If our judgement about narrative integration required knowledge of whether Karen accepts her motive as her own, then it would not be independent of the "generic" notion of "G-identification".

[46] Of course, I have argued that G-identification is not sufficient to establish that the motive is *truly Karen's own*. For example, suppose a cause of her action of ordering chocolate cake is that she earlier worked on a complex logical problem. If Karen knew that this was a cause, she might not G-identify with her motive. In this case, the motive is not truly her own, in my view.

[47] She might never before have "indulged herself" by ordering an unhealthy dessert.

[48] In this discussion of Fatima's case I am ignoring the possibility that she might embrace the motive if she knew the action's *causes*.

[49] I take it that this is the sort of action that Doris has in mind in his discussion of identification with an unconscious desire to please one's father, though he does not give an example of such an action (2002:141-142).

[50] In my terms this desire (or more strictly, the *onset* of this desire) is a subverting cause of Fatima's action. I didn't include an example of this kind among my five key cases because I'm not aware of research which demonstrates this kind of effect. (If such effects do indeed occur, then they lend support to my claim that there are many everyday actions with subverting causes, and so too to my sceptical argument in chapter 7).

identification need not excuse her from moral responsibility for the action which issues from that desire, because she may bear TMR for it.

Fatima's unconscious desire has influenced many of her previous actions, including several which are harmful to her colleagues. I'll consider two possible versions of this history. In the first version, she is aware that several of her previous similar actions have been detrimental to her colleagues. She had the opportunity to reflect on the effects of her actions, and whether she regretted or was satisfied with those effects. After reflecting, she was satisfied with those effects, and she decided not to take steps to change her future behaviour. This decision was fully her own, and she was directly morally responsible for it. To this decision we can *trace* Fatima's moral responsibility for her current action – indiscreetly revealing a secret about her colleague. Fatima bears TMR for revealing the secret, because she was directly morally responsible for an earlier action (the decision to do nothing to change her future behaviour), and the later action was reasonably foreseeable by her at that time.[51] There is an obvious parallel with Vargas's example case of Jeff the jerk (discussed in section 1.2). I think the same arguments apply, and Fatima's responsibility can be traced to her earlier decision.[52] In this first version of Fatima's history, we have no need for Doris's notion of narrative integration: the established principle of tracing gives the conclusion which reflects our intuitions and reactive practices.

In the second version of Fatima's history, she reflected on the effects of her previous actions and was not satisfied with them. She embarked on a course of therapy with the intention of changing some of her future behaviour. However, she has so far been unable to make significant changes to that behaviour. She also remains unaware of the unconscious motive and would continue to disavow it if made aware of it. Here I think the only plausible verdict is that Fatima *is not identified* with her unconscious motive. She might still be morally responsible for revealing the secret, if responsibility can be traced to some earlier point in her history. But, contrary to Doris's NI condition, she is not identified with the motive despite the fact that it admits of narrative integration.

Thus on the first version of Fatima's history narrative integration is superfluous; on the second it yields an implausible verdict. Tracing captures the way in which Fatima can be morally responsible for her action, in a more plausible manner than the claim that she is identified with a motive which she would genuinely disavow, if she knew of its existence.

Problems also arise, I think, from Doris's discussion of the hard-charging stockbroker (whom I called Steve). Doris suggests that Steve can be held morally responsible for walking past a person in distress, because he "embrace[s] having the sort of packed schedule that induces haste".[53] This is despite the fact that Steve "fails to identify with the callousness that resulted from haste". Steve "reject[s] his many hasty omissions, but wholeheartedly endorse[s] the way of life that leads to them". Doris seems to think that Steve is not G-identified with his determinative motive, and that this is a weakness in the generic account of identification, which can be rectified by using the narrative integration condition instead. I'm not convinced that this analysis is correct. It's not clear that Steve fails to embrace (or regard as fully his own) the determinative motive which leads him to walk past a person in distress, which is presumably a motive which concerns meeting other commitments. If indeed Steve "wholeheartedly endorses the way of life" with which that motive is surely consistent, then it

[51] An incompatibilist might want to add the condition that it must have been possible for Fatima to change her future behaviour. I am sympathetic to this claim, but I set it aside so that my arguments will apply to compatibilists who would reject it.

[52] Here we may treat the earlier decision as an action, or as an omission (since it's a decision to do nothing). I think the point succeeds either way.

[53] All the quotes in this paragraph are from Doris 2002:144.

seems more plausible to say that Steve does embrace that motive.[54] In that case, he *is* G-identified with the motive, and narrative integration is once again superfluous.

I'll set aside that point, though, for the sake of argument. I'll assume that Steve does not G-identify with his effective desire to walk past the person in distress – he does not embrace it or regard it as his own. But Steve does meet the narrative integration (NI) condition. I think the case is now analogous to the first version of Fatima's history I envisioned above: Steve is not DMR for his action, but he is very likely to be TMR for it. It's very likely that he has had the opportunity to reflect on the effects of his habitually hasty behaviour, and to decide whether to take steps to change his future behaviour. Again, we have no need for narrative integration to reach this result.

Now suppose instead that when Steve walks past someone in distress his action has *a situationist cause*: suppose that he is in the same position as Samantha and the participants in the 'Good Samaritan' experiment (section 3.2.1). He walks past because he was told that he is running late, but would have stopped to help in the absence of that event. I take it that Doris views Steve as identified with his effective desire and morally responsible for his action, because the action reflects his operative priorities or evaluative orientation, as revealed in a suitable narrative.[55]

I think there is a serious problem with this view. It seems to be at odds with Doris's own (persuasive) arguments about the powerful influence of situationist causes. The disturbing feature of situationist causes is that agents' subsequent actions are not predictable on the basis of their character or past habitual behaviour. The research reveals that features *of the situation* have a surprisingly marked effect on agents' actions. They are surprising precisely because they *do not* fit easily into agents' narrative histories. So in the presence of situationist causes, a person's narrative history is not a suitable guide to whether he is identified with an effective desire. Two kinds of error could occur, if we use narrative history as the condition of identification.

First, when there is a mesh between the action and the agent's narrative history, the mesh might be *merely coincidental*. When Steve walks past the person in distress after being told that he is late for his presentation, this action meshes with his narrative history. But in the original experiment, 90% of subjects (who were seminarians rather than hard-charging brokers) did the same thing. Given these facts, we cannot conclude that Steve's action reflects his operative priorities or evaluative orientation; it may instead simply reflect the power of that situational factor. If in fact the situational factor was a cause of his action, and Steve would not embrace his motive upon learning of the action's causes – in other words, if my CFG condition is not met – then I submit that the motive is not fully his own.

In the second kind of error, we might excuse from responsibility someone who is not habitually "hard-charging", on the basis that her action does not mesh with her narrative. But on this particular occasion she might be identified in the sense that my CFG condition captures. That is, she might embrace her effective desire to walk past the person in distress, even if she were informed that being told she was late was a cause. In my view, it's much more plausible in that case to say that the desire is fully her own, and to hold her responsible for the action.

Thus I think narrative integration is flawed as a guide to moral responsibility when there are situationist causes of actions. The same is true, for the same reasons, for other types of subverting cause. The counterexamples in the last two paragraphs show that Doris' NI

[54] Doris's phrasing is peculiar when he says that Steve "fails to identify *with the callousness* that resulted from haste" (my italics), rather than failing to identify *with a motive*. This may be a sign that the case cannot support the weight placed upon it.

[55] Doris's remarks in this part of his discussion (2002:144) suggest that this would be his conclusion, though, unfortunately, he does not mention any specific examples in which the "hard-charging broker" would bear responsibility despite "situationally induced motives".

condition is neither necessary nor sufficient for a desire being fully one's own, and hence neither necessary nor sufficient for direct moral responsibility.

5.1.2. Doris's conclusions on situational factors and responsibility

Doris reaches several conclusions about responsibility in the light of his discussion of situational influences on actions. I confess that I find the conclusions confusing; it seems to me that some conclusions conflict with others, and I don't see how the conflicts are to be resolved. In the following discussion I'll highlight some of the difficulties.

It's not clear to me whether Doris thinks that situational effects could ever excuse an agent from moral responsibility for his action. Though Doris's account of responsibility is clearly intended as a response to the threat of situationism, his discussion of narrative integration includes only one (very brief) example of agents performing actions unknowingly influenced by situational pressures: of the subjects in the Milgram experiments who decide to administer strong electric shocks as instructed, Doris thinks we can easily take either view – that they are responsible, or that they are not. He comments "I suspect that … ambivalence is the appropriate judgment – or abstention from judgment".[56]

In his 'Conclusion' section to the chapter on responsibility, Doris considers this question:

> Is it the case that those unfamiliar with the tradition of experimental social psychology suffer exculpating ignorance, insofar as they lack our keen awareness of situational danger? This result is to be resisted, lest thinking on responsibility become implausibly tender-minded. It is quite true that some situational factors, such as group size effects, are unknown to those lacking familiarity with a specialized literature, and here there may be surprising instances of exculpating ignorance. But ignorance of a bit of psychological theory does not excuse ignorance of an independent and ethically relevant fact.[57]

To support this last point Doris draws on a highly-charged real-life example of a woman who was stabbed repeatedly, but some 38 bystanders apparently did not help or even call the police until it was too late.[58] It seems likely that situational factors – such as the behaviour of the other bystanders as a group – were causes of some bystanders' actions. Of these people Doris says:

> the Genovese bystanders were painfully aware of a person in desperate need of help, even if they were unaware of situational factors implicated in their failure to provide it. Excusable ignorance in one area does not necessarily excuse negligence in another.[59]

On the other hand, Doris suggests that "encountering situationism facilitates increased personal responsibility".[60] Once aware of such factors we can – and we *should* – avoid situations in which those factors are likely to influence our actions.[61] For example Alf, once aware that his flirting over dinner might lead to a later act of akratic infidelity, should avoid flirting. This pushes the "locus of responsibility" to an earlier point in the process – in this case, to the time of the flirting rather than the time of the infidelity. Indeed, Doris even

[56] Doris 2002:145. See section 3.6.1 for a description of the Milgram experiments.
[57] Doris 2002:153.
[58] The victim was Catherine ("Kitty") Genovese, murdered at night in Chicago in 1963. Earlier (2002:28-29), Doris describes key details of the incident, citing a book written by a newspaper editor.
[59] Doris 2002:153.
[60] Doris 2002:129.
[61] Doris 2002:147.

suggests that we have a "cognitive duty" to attend to possible situational factors, once we are aware of them.[62] Doris worries that:

> To some, my recommendations may seem more a recipe for treating oneself as a pet than for deliberating as a responsible person.[63]

But he concludes that:

> Better understanding the determinants of behavior facilitates a process of self-manipulation that allows people to take a more active and responsible role in our own lives.[64]

There seem to be two main conclusions here. Knowledge of situational factors facilitates increased responsibility. But a lack of such knowledge is not exculpating. It seems to me that there is a tension between Doris's two conclusions. If knowledge of situational factors facilitates increased responsibility, then the corollary surely follows that people who lack that knowledge must bear less responsibility. But Doris's phrasing in these passages is rather open-ended. He speaks of taking a "more responsible role in our lives", but he gives no examples of agents who either do or do not bear responsibility for specific actions which have situationist causes. In the long quote above he urges us to resist "exculpating ignorance" – but it's not clear whether he means that the bystanders to the woman's stabbing should be held *responsible* or *blameworthy*.[65] He doesn't discuss exactly what actions or omissions he takes them to be responsible or blameworthy for; nor does he give a detailed argument for his conclusion in that particular case. Nor does he attribute responsibility (or blameworthiness) in any other specific cases of actions which have situational factors as causes.

I'll consider two such actions, and argue that Doris's conclusions on each one are misguided.

Doris argues (plausibly) that Alf bears responsibility when he flirts over dinner and subsequently commits an act of infidelity. No specialised knowledge of social psychology research is required in this case: the situational pressures are all too predictable. Doris proposes (implausibly) that Alf is responsible for the flirting over dinner, but not for the akratic infidelity. (As I argued in the previous section, Doris's proposal would involve an unnecessary change to our reactive practices. We intuitively hold Alf responsible *for the act of infidelity*, and we do so by employing the tracing principle.)

But it's not clear to me how Alf's knowledge about the situational pressures is supposed to contribute positively to his bearing responsibility for his actions.[66] Suppose Albert is in the same situation as Alf, but is extremely naïve. Albert spent his formative years in a monastery and got married very soon after leaving; the person with whom he flirts over dinner is only the second woman he has got to know at all well. Albert does not realise that his flirting might lead to a later akratic infidelity. For which action is Albert responsible? Doris presumably holds Albert responsible for his flirting, and not for the akratic infidelity which follows. But this is the same verdict as Doris gives about Alf, who *did* know about the situational effects he might be subjecting himself to when he flirted. I don't see how Alf's extra knowledge leads to "increased" responsibility, on Doris's account. (On a standard view, in contrast, Alf bears more responsibility than Albert – in one important sense at least. Both are directly

[62] Doris 2002:148.
[63] Doris 2002:149.
[64] Doris 2002:153.
[65] According to many philosophers, one can be morally responsible without being blameworthy, (see also section 1.1.3). Doris doesn't state his view on this, as far as I'm aware.
[66] I take it that if knowledge about situationist effects "facilitates increased personal responsibility", then it must in some way contribute positively to bearing responsibility for specific actions.

morally responsible for their flirting, but only Alf bears traced moral responsibility for the later infidelity.)

Meanwhile, what will Doris say about Samantha's responsibility for her action of walking past the slumped man, if she is not aware of situationist research?[67] On the most natural reading of the long quotation above, I take it that Doris would hold Samantha responsible for her action, on the grounds that "ignorance of a bit of psychological theory does not excuse ignorance of an independent and ethically relevant fact". But Doris's phrasing in this important point is curious, and he does not spell out exactly what he means by it. There are at least three kinds of ignorance which Samantha might exhibit here. First, she *is* ignorant of the relevant situationist psychological theory. Second, she might be ignorant of the relevant moral principle. In this case we might say that the relevant principle is that people have a moral obligation to help others who may be in distress. Third, she might be ignorant of empirical facts in this particular situation. For example, she might not recognise that the slumped man is showing signs of distress, or that the audience for her presentation wouldn't be inconvenienced by her late arrival.

I'll suggest two ways to interpret Doris's point. The first is that he is appealing to an axiom about responsibility: ignorance *of a moral principle* does not excuse someone from responsibility if she flouts that principle. If Samantha were ignorant of the moral principle that people have a moral obligation to help others who may be in distress, this would not excuse her for failing to help someone showing signs of distress. But there's no reason to think that Samantha is ignorant of that principle. Rather, as I'll discuss in more detail below, it may be that Samantha is ignorant *of certain empirical facts* which obtain in this situation.

On the second interpretation, Doris is making the point that agents cannot be exempted from responsibility if they *deliberately ignore* an empirical fact which has moral significance.[68] At first glance this seems an unlikely interpretation, since Doris's phrasing suggests that he is talking about a passive *ignorance* rather than an active *deliberate ignoring*. But it does lead to a more intuitively plausible line of argument, which is worth discussing. Because the situational effect in the 'Good Samaritan' case is so extreme and so surprising, it seems natural to assume that all the agents who fail to help the slumped man must have deliberately ignored his signs of distress. But it turns out that this is another instance in which our everyday model of action explanation is too simplistic.

As I discussed in section 3.8.5, situational factors sometimes distort an agent's assessment of her circumstances. In their discussion of the 'Good Samaritan' study, experimenters Darley and Batson noted:

> According to the reflections of some of the subjects, it would be inaccurate to say that they realized the victim's possible distress, then chose to ignore it; instead, because of the time pressures, they did not perceive the scene in the alley as an occasion for an ethical decision.[69]

Suppose that Samantha simply does not recognise the slumped man's possible distress. We cannot then hold her morally responsible on the grounds that she *ignored* "an ethically relevant fact". She did not ignore the man's possible distress; she simply did not recognise it.

Now, some subjects in the 'Good Samaritan' study did seem to notice and assess the man's plight before walking past without helping. Of these, some showed signs of being conflicted about that action, but it appears that they did not assess the man's possible need for

[67] Samantha walks past a slumped man on her way to give a presentation about the parable of the 'Good Samaritan' (see section 3.2.1).
[68] I addressed a similar point in the form of an objection to my proposal, in section 4.3.2.
[69] Darley and Batson 1973:108.

help as being *greater* than the obligation to give their presentations on time.[70] It's surely implausible to think that the vast majority of subjects recognised that the greater moral obligation was to help the man, but pressed on regardless.[71] If that is correct, then it would be false to say of many subjects that they *ignored* "an ethically relevant fact", in the sense that they were aware that the moral obligation to help the man was greater than the moral obligation to give their presentations on time, and yet they decided not to act on it.

On either interpretation, then, I think Doris's point fails to hit home. Meanwhile Doris himself seems to overlook a very important feature of cases like Samantha's. It seems very likely indeed that *many* of the subjects in the 'Good Samaritan' experiment assessed the slumped man's plight differently than they would have done without being told of a time pressure.

In my view, Samantha is excused from *direct* moral responsibility because the motive on which she acts is not fully her own: she would not regard it as fully her own, if she knew all of the causes of her action. Samantha's assessment of the situation is distorted by the situational factor – in this case, being told that she is late – so that she does not recognise the greater moral obligation before her. What excuses Samantha from *traced* moral responsibility is that she is ignorant of the effect that the situational factor may have on her future assessment. It is not reasonably foreseeable by her that she might act without recognising the greater moral obligation, and she takes no precautions against doing so.

What of the specific case in which Doris does attribute responsibility – the 38 witnesses to murder who "were painfully aware of a person in desperate need of help", as Doris describes them? It's difficult and potentially misleading to draw conclusions about social psychology from journalists' accounts of a single event. A recent investigation into the circumstances throws considerable doubt on the key features that Doris appeals to: "there is no evidence for the presence of 38 witnesses, or that witnesses observed the murder, or that witnesses remained inactive".[72] So it's far from clear that there were 38 people on that night who deliberately ignored "an ethically relevant fact", and who should be held morally responsible on those grounds. Furthermore, even if there were, it would not follow that Samantha is morally responsible for *her* action on the same grounds. Experimental research on the subject of bystander intervention reveals that the most natural explanation of Samantha's action is too simplistic.

5.1.3. Summary of my conclusions on Doris's account

Doris's account of responsibility for actions, unlike Frankfurt's, recognises the relevance of events which occurred *before* the current action. The notion of narrative integration is a potential substitute for the tracing principle, since it takes account of the agent's personal history including his previous actions. At the same time it addresses the challenge posed by subverting causes.

But I argued that there are two respects in which tracing is simply more plausible than narrative integration. It's much more plausible to say that Alf bears moral responsibility *for his akratic act of adultery*, than merely for his earlier flirting from which the adultery resulted. Secondly, Fatima's unconscious desire to please her father *does not reveal her true*

[70] Darley and Batson 1973:108.

[71] In the experiment, only 10% of subjects told they were running late offered to help, compared with 63% of those told they were early (see section 3.2.1).

[72] Manning, Levine and Collins (2007:555). In Doris's defence, the Kitty Genovese case has been included in psychology textbooks for several decades as an example of the power of social group effects, though it now seems to be a very dubious example.

self, if she would disavow it upon learning about it. Tracing is a more plausible means by which she can still be held morally responsible for actions which issue from it.

But even if we set aside these arguments that narrative integration is much less plausible than tracing, there remain other serious problems with Doris's account. It seems that there is circularity in Doris's use of narrative integration as a condition of moral responsibility: it's far from clear that our judgements about whether an agent's desire to act admits of narrative integration are *independent* of our judgements about whether the agent is morally responsible for an action. Doris's appeal to policies does not help him to evade the circularity, and at the same time introduces further problems. Karen's akratic ordering of chocolate cake contradicts both her policy and an established narrative which expresses her priorities, and yet we will still hold her morally responsible for it. When Sidney breaks a new policy to stop smoking after many years, and that policy itself contradicts an established narrative, it is not clear what verdict will be reached on Doris's account.

Meanwhile Doris's treatment of actions with subverting situationist causes is problematic. He claims that knowledge of situational factors facilitates increased responsibility, but a lack of such knowledge is not exculpating. These claims seem to be in tension with one another.

Doris would hold the hurried broker Steve morally responsible for walking past a slumped man after being told that he is running late for a presentation, on the grounds that such an action meshes well with an established narrative. But this seems to undermine Doris's earlier arguments that situationist causes often issue in actions which contradict the agent's history and habitual behaviour. I think implementing Doris's view would risk holding Steve morally responsible when the action is not fully his own – if he would not regard it as fully his own, upon learning of its causes. Equally, Doris's line would excuse someone in the same situation whose walking past the bystander *was* fully his own, but did not mesh with a narrative. Therefore I think Doris' NI condition is neither necessary nor sufficient for direct moral responsibility.

Also troubling is Doris's insistence that exculpation of agents who lack knowledge of situationist causes "is to be resisted". He attempts to judge responsibility in only one case of a well-researched action with a situationist cause: when Milgram subjects follow instructions to administer shocks, he suggests that *abstention* from judgement is appropriate. The sole example in which he does clearly judge responsibility is not a researched case at all, and serious doubts have been raised about the facts that are most relevant to his judgement. Doris's analysis doesn't make adequate allowance for the distorting effect of situational factors on agents' *assessments* of the situations they face.

In the case of Alf, whose flirting leads on to an akratic infidelity, Doris proposes a change to our reactive attitudes, so that we hold Alf responsible for the flirting but not the infidelity. I think this is both unnecessary and unlikely to happen. I think instead that there may be a different kind of change, as we come to understand subverting causes more thoroughly, and as knowledge about them spreads through the general population. Rather than shifting the object of our reactive attitudes from one action to another, we may come to see that our reactive attitudes are misplaced for many actions. I'll have more to say about this in chapter 7.

5.2. EDDY NAHMIAS

Another philosopher who has engaged deeply with situationist research is Eddy Nahmias.[73] Nahmias aims to give an analysis of free will which "refines" Frankfurt's by adding requirements concerning the agent's knowledge about her own desires.[74]

Nahmias is primarily focused on analysing free will, rather than moral responsibility. He distinguishes between *possession* and *exercise* of free will, both of which, in his view, admit of degrees. Possession of free will is necessary for its exercise. Exercise of free will in a particular action is necessary for moral responsibility for that action, which also admits of degrees.[75]

Nahmias proposes a "Knowledge Condition", necessary for possession of free will, which involves three cognitive abilities.

> First, we must have introspective access to at least some of our most important desires; we must recognize what desires we have, how they conflict, and how they might motivate us to act. Second, we must be able to care about how we act and how we are motivated to act; we must be able to identify ourselves with some of our desires and not others. We must know what we really want. Third, we must be able to become motivated by those desires we identify with; this will involve a form of control over our desires that often requires knowledge of their influence on us.[76]

The kind of access to our own desires that we require does not involve continuous introspection.

> It would be unrealistic for a theory of free will to require that we self-consciously consider our desires every time we make a choice … Nonetheless, awareness and reflection are not irrelevant; identification does require that the agent has the *ability* to introspect on her motivations and reflect on her attitudes towards them.[77]

Nahmias sees the Knowledge condition as necessary both for possession of free will and for identification; while identification is necessary for the exercise of free will and hence for moral responsibility for a particular action. For identification itself,

> it is not necessary that we are in fact aware of what moves us to act – only that we would identify with our motivations *if* we became aware of them.[78]

I'll call this the "Counterfactual Awareness" condition: *a determinative motive is truly the agent's own only if she would identify with it if she became aware of it*.[79]

---

[73] Nahmias's Ph.D. dissertation (2001) covers the situationist literature in much more detail than I have done. He does not discuss other types of subverting cause.

[74] Nahmias makes many criticisms of Frankfurt's theory, some of which are unrelated to situationist research findings.

[75] Nahmias 2001:76. Nahmias does not think that the exercise of free will is sufficient for moral responsibility, since normative competence of the agent is also necessary (2001:148ff). In my discussion, I will ignore the complication that these concepts may admit of degrees, since I don't think it makes any significant difference to my arguments.

[76] Nahmias 2001:4.

[77] Nahmias 2001:108-109.

[78] Nahmias 2001:184. Nahmias makes some very interesting observations about the kind of agential history required for identification, which I won't discuss here.

[79] This condition can be made independent of any particular account of identification. It can be applied, for example, using the generic definition outlined by Doris: to identify with one's motive is to embrace it or regard it as fully one's own. (I called this "G-identification" in section 5.1). I assume that the counterfactual identification would occur *at the time of acting*.

Nahmias therefore concludes that situationist findings do threaten the exercise of free will (and hence moral responsibility for actions).

> So, subjects in the [situationist] experiments are usually not acting of their own free will. Even *if* they became aware of their motivations or the process by which they acquired them, they would usually not identify with them. They would feel constrained by them and try to overcome their influence.[80]

Nahmias concludes his discussion of the threat to free will posed by situationist experiments with a call for further research. He seems to have an open-minded view of what that further research will bring.

> While such investigations may indicate limits on our freedom, they also have the potential to increase our knowledge of ourselves in ways that augment our free will.[81]

5.2.1. Criticisms of Nahmias's account

I'll focus my critical discussion on Nahmias's Knowledge condition and the Counterfactual Awareness condition. I'll argue that these conditions are not jointly sufficient for moral responsibility. I'll then argue that neither is necessary for moral responsibility.

I'll begin by considering whether the Knowledge condition and the Counterfactual Awareness condition are jointly sufficient for moral responsibility. For this purpose I will set aside the other conditions which Nahmias holds to be necessary, principally those concerning normative competence and the exercise of free will; I will assume that those conditions are met.

I'll take the 'Good Samaritan' case as my example. When Samantha walks past the slumped man to deliver her presentation, her determinative motive is wanting to reach the room where she will be speaking. Other relevant desires may include the desire to please or help the person who asked her to give the presentation, the desire to appear competent or reliable, the desire to give a successful talk, and so on. She may or may not have a desire to help the slumped man.[82] If she does have that desire, it is less strongly motivating than the combination of the other desires. I see no need to assume that Samantha is not consciously aware of any of these desires. Much less plausible still is the notion that Samantha *cannot* introspect on them, and reflect on her attitudes towards them. So it seems that Nahmias's Knowledge condition is met for this action.

Meanwhile, I see no reason to think that the Counterfactual Awareness condition is not met. That condition requires that Samantha *would have* identified with her determinative motive, had she become aware of it. As I've explained, it's very plausible to think that Samantha *was* aware of all the desires which moved her to act. Furthermore, it seems very likely that Samantha did identify with her determinative motive. She did not know, of course, about the influence of the subverting cause. So it's very plausible to think that both the Knowledge and the Counterfactual Awareness conditions are met, in Samantha's action.

And yet, even if that is the case, I think Nahmias is right to conclude that the effective desire is not truly Samantha's own. The problem is not that Samantha lacks knowledge *of her desires* as they are. The problem is that she lacks knowledge of *how her desires have been influenced* by the subverting cause, and hence how her action has been affected. The subverting cause is that she was told that she was running late. If she became aware that she

[80] Nahmias 2001:185.
[81] Nahmias 2001:239.
[82] As discussed in section 5.1.2, it may be that she does not assess him as requiring any assistance.

would have acted otherwise without this cause, she would not regard her effective desire and her action as truly her own. That is, she would not meet my CFG condition of direct moral responsibility.[83] With knowledge of how such causes can sometimes operate, gleaned perhaps from studying situationist psychology research, Samantha might instead have stopped to help the slumped man. That action would have been truly her own, and she would have been directly morally responsible for it.[84] But having the ability to introspect on her desires would not have made any difference here. What would have helped, instead, is knowledge of how subverting causes can operate.[85]

Knowledge of how subverting causes operate might also allow agents to bear *traced* moral responsibility for their actions. For example, if Gina had known that giving up on a problem might result from resisting cookies (section 3.2.5), she might be TMR for doing so. By eating the cookies (or perhaps by not leaving them in full view), she could have avoided giving up on the problem. But again what prevents her from bearing responsibility in the actual situation is not lack of knowledge *of her desires*, or lack of an ability to know them. Gina is all too aware that she desires to eat cookies, and that she also desires not to eat cookies (i.e. she wants to resist eating them). She may have related desires about losing weight or saving the cookies for someone else. Later, she desires to complete the problem, and also has a conflicting desire to give up on it. It is not because of a lack of introspective access to these desires that the action of giving up on the problem is not truly her own. The important knowledge she lacks is about the subverting effect of the resisted cookies.

I conclude from this discussion that the Knowledge condition and the Counterfactual Awareness condition are not jointly sufficient for moral responsibility. An agent may have introspective access to all the relevant desires at the time of acting. It may be true that she was aware of, and did identify with the effective desire at the time of acting. And yet, it may also be true that the desire is not truly her own, because she would not identify with it if she knew all the causes of her action. If the effective desire is not truly her own, then she is not directly morally responsible for the action.[86]

I'll now argue that the Knowledge condition is not necessary for moral responsibility. I take it that one of the central claims made in some schools of psychotherapy is that we simply cannot gain introspective access to certain 'deep-seated' unconscious desires. Take for example Doris's case of Fatima, who has an unconscious desire to please her father (see section 5.1 above). Suppose further that Fatima cannot gain introspective access to that desire. I think it's plausible to say that Fatima can be morally responsible for an action which issues from that desire, such as revealing a secret about a colleague, despite her being quite unaware of the desire's existence. It might be that she would meet my CFG condition: it might be that she would embrace her effective desire or regard it as fully her own if she knew all of the action's causes. In that case she would be DMR for her action. Meanwhile if Fatima is aware that her previous similar actions have caused problems for her colleagues, and she has decided to do nothing to change her behaviour, she can bear *traced* moral responsibility for this latest action.[87] In either case, I think holding Fatima morally responsible is much more plausible than excusing her, as Nahmias apparently would do, because the Knowledge condition is not met.

---

[83] I discussed the CFG condition in section 5.1.1 above. It requires that: the agent *would embrace her determinative motive or regard it as fully her own*, if she knew all the proximate and relevant *causes* of the action.

[84] In that scenario, being told that she was late would not have been a subverting cause.

[85] It may be that knowledge of how subverting causes can operate is necessary for *possession of free will*. That is not a conclusion I am arguing for, so I won't follow up this line of thought.

[86] This is an identificationist axiom, equivalent to statement (4) with which I defined Frankfurt's position in section 2.2.

[87] I argued for this conclusion in section 5.1.1 above.

An objector might deny that my presentation of Fatima's case here is realistic; he would deny that there are important unconscious desires to which agents cannot gain introspective access. This is indeed a controversial issue in psychology. But even if the case is only hypothetical, I think the conclusion is clear enough: an agent could bear moral responsibility for an action *even if* it issued from an unconscious desire to which she could not gain introspective access.

My arguments here may not be convincing. But it doesn't matter to my overall position if Nahmias is correct on this point. If the Knowledge condition is indeed necessary for moral responsibility, the result is fewer actions for which it is possible for an agent to be morally responsible. Since my conclusions are in the end sceptical about the prevalence of moral responsibility, I am not troubled by this possibility.

Doris claims that an agent *can* be identified with a motive which she would disavow after learning of it. His example is that of an agent with an unconscious desire which she would disavow. I argued that, in such a case, it is more plausible to say that the agent's desire is *not* fully her own if she would disavow it. And yet I do not think Nahmias's Counterfactual Awareness condition is necessary for direct moral responsibility. Instead, I think my proposed CFG condition is necessary for DMR, and does not need to be supplemented by the Counterfactual Awareness condition.

The Counterfactual Awareness condition gives a different verdict from a standard or "generic" condition of identification when there are determinative motives *of which the agent is unaware*.[88] In those situations, I think my CFG condition will give the same verdict as the Counterfactual Awareness condition. For example, take the case of Fatima, whose unconscious desire to please her father endorses her effective desire to reveal a secret about a colleague.[89] It's likely that, if she knew of the unconscious desire, she would disavow it. She would not meet the Counterfactual Awareness condition. I think that, if instead she knew all the *causes* of her action, she would again disavow the effective desire. These causes would be subverting, and she would not meet my CFG condition. In other words, the results would be equivalent in these two counterfactual scenarios, and we would conclude, using either condition, that Fatima is not directly morally responsible for her action.

What causes might Fatima become aware of, which would lead her to disavow the effective desire? I think there are two potential kinds. The first kind would include incidents from Fatima's past, such as occasions on which her father perhaps scolded her for her lack of competitiveness, or rewarded her for breaking a confidence. Fatima does not realise that these incidents are causes of her current action; if she became aware of that fact, she would disavow the desires which led her to act. The second kind of cause I have in mind includes events much more proximate to the action, such as Fatima's looking at a photograph of her father immediately before acting, or her boss behaving in a way that reminds her (perhaps even unconsciously) of her father. If Fatima were made aware that these events were causes of her action – and that she would have acted otherwise without their presence – she would disavow her effective desire.

I conclude that, given that the CFG condition is necessary for direct moral responsibility, the Counterfactual Awareness condition is not also necessary. Of course, by giving these limited examples I have not made a very strong case that for that conclusion. But I'm content to say no more, since my main objective is to show that the CFG condition *is* necessary. Furthermore, if there are everyday actions for which the Counterfactual Awareness condition indicates a *more* sceptical verdict about the prevalence of moral responsibility than

[88] The "generic" identification condition discussed by Doris was: to identify with one's determinative motive is to embrace it or regard it as "fully one's own".

[89] My proposed CFG condition was: The agent would embrace her determinative motive or regard it as fully her own, if she knew all the proximate and relevant causes of the action. See section 5.1.1.

the CFG condition, my secondary objective in chapter 7 below will actually be helped. For there I will argue that there are many everyday actions for which we mistakenly hold agents morally responsible.

5.3. CONCLUSIONS

In chapter 3 I argued that Frankfurt's identification condition is not sufficient for direct moral responsibility. An action with subverting causes is not truly the agent's own, and hence the agent cannot be directly morally responsible for it. In some actions with subverting causes, though, Frankfurt's identification condition is met. I proposed an alternative identification condition which I believe is both necessary for direct moral responsibility and acceptable to followers of Frankfurt.

In this chapter I've considered two identificationist accounts developed specifically in response to the threat to moral responsibility posed by actions with situationist causes – those of Doris and Nahmias.

I argued that Doris' narrative integration NI condition is neither necessary nor sufficient for DMR. I proposed that the following condition, which incorporates a very generic notion of identification, is necessary instead for direct moral responsibility:

> *CFG* The agent *would embrace her determinative motive or regard it as fully her own*, if she knew all the proximate and relevant *causes* of the action.

Against Nahmias, I argued that his Knowledge condition and Counterfactual Awareness condition are not jointly sufficient for direct moral responsibility. I also argued that neither is necessary for DMR.

In the next chapter I'll address a different kind of compatibilist theory, which is not primarily identificationist – that of Fischer and Ravizza.

6. FISCHER AND RAVIZZA

John Martin Fischer's theory of moral responsibility has developed gradually over several years. I will concentrate on the position discussed in his book written with Mark Ravizza, *Responsibility and Control*, and subsequent papers.[1] Most of the time I'll refer to it as "the Fischer-Ravizza account", but it will be clear from references when I'm discussing a contribution made by Fischer alone.

A central element of the Fischer-Ravizza theory is that an ability to respond to reasons is necessary for moral responsibility. To that they add that the agent must "take ownership" of his actions. By combining these elements in the right way, Fischer and Ravizza believe they can give an analysis of the appropriate form of *control* required for moral responsibility.

Fischer and Ravizza's complex theory contains many insightful innovations, and has already become extremely influential: it has recently been described as "the gold standard for cutting edge defenses of compatibilism".[2] I'll begin (section 6.1) with an overview of the theory, followed in sections 6.2 to 6.5 by a description of its key components. In section 6.6 I'll discuss some important criticisms raised by other philosophers, and I'll conclude from those that the Fischer-Ravizza theory contains several flaws.

I'll set aside those criticisms to build my main argument against the theory, in section 6.7. By examining actions with subverting causes, I'll argue that the conditions set out by Fischer and Ravizza are not sufficient for direct moral responsibility.[3]

6.1. OVERVIEW OF THE THEORY

Fischer and Ravizza endorse Harry Frankfurt's attack on the 'Principle of Alternate Possibilities' (PAP). According to PAP: "a person is morally responsible for what he has done only if he could have done otherwise".[4] Frankfurt denies PAP because he thinks that agents in certain kinds of case (now widely called 'Frankfurt-style' cases) are morally responsible despite lacking alternate possibilities.

Here's one Frankfurt-style case.[5] Jones is about to cast a vote in an election. He does not know that Black, a well-meaning liberal neurosurgeon, has placed a microchip in his brain. Using the microchip, Black can monitor Jones's brain and cause him to act as she chooses. In fact, Jones votes Democrat in the election, and Black does not intervene. But if Jones had been about to vote Republican, Black would have activated the chip and ensured that Jones voted Democrat instead. Frankfurt holds – and Fischer and Ravizza agree – that Jones is morally responsible for his action, despite the fact that he had no access to alternate possibilities (i.e. despite the fact that he could not have voted Republican).[6]

It seems plausible to say that Jones exercises *some* kind of control over his actual action when there is no intervention by Black. Fischer and Ravizza aim to specify the kind of control which is necessary for moral responsibility. They distinguish two kinds of control: regulative control, and guidance control. For *regulative* control of an action, the agent must have alternate possibilities. Jones lacks regulative control of his voting action, because he has no

[1] The book covers moral responsibility for *actions*, *omissions* and the *consequences* of actions and omissions. The scope of my enquiry is limited to moral responsibility for actions, and I'll restrict my discussion to that.
[2] McKenna 2004.
[3] Direct moral responsibility (DMR) is anchored in facts about the action and the agent at the time of acting. In contrast, traced moral responsibility (TMR) can be "traced back" to facts about an earlier action (or omission) and about the agent at that earlier time. See section 1.2.
[4] Frankfurt 1969:1.
[5] This case is adapted from Fischer 2007:58.
[6] There are many objections to Frankfurt's argument, but they are not relevant to my discussion here. I will assume for the sake of argument that Fischer and Ravizza are correct to accept Frankfurt's conclusion. For an overview of the large and complex debate surrounding Frankfurt's argument, see for example Fischer 2002b.

alternate possibilities. Therefore, Fischer and Ravizza conclude that regulative control is not necessary for moral responsibility.[7]

Jones is morally responsible for his voting action, according to Fischer and Ravizza, because he has *guidance* control of the action: "guidance control is the freedom-relevant condition necessary and sufficient for moral responsibility".[8]

There are three key components in the Fischer-Ravizza account of guidance control: a *mechanism* which "issues in" an action; the agent's *ownership* of that mechanism; and the mechanism's being appropriately *reasons-responsive*. When an action issues from the agent's own appropriately reasons-responsive mechanism, the agent is said to have guidance control over the action. I'll describe these three components in the next three sections.

6.2. "THE MECHANISM"

The notion of "the mechanism issuing in an action" plays a crucial role in the Fischer-Ravizza account, but it is not precisely defined. No examples of individual mechanisms are given. The authors give perhaps their clearest indication of what a mechanism actually is when they explain that it is not "like a mechanical object":

> although we employ the term "mechanism" we do *not* mean to point to anything over and above the process that leads to the relevant upshot; instead of talking about the mechanism … we could instead talk about the process that leads to the action, or the "way the action comes about".[9]

The authors claim that mechanisms can be grouped into "*kinds of* mechanism". They cite as kinds of mechanism "practical reasoning, non-reflective habits, and so forth".[10] Many paradigmatic actions issue from "practical reasoning". An example of an action issuing from "non-reflective habits" is turning off the freeway at the exit that one uses every day, without conscious deliberation.[11]

The notion of a *kind* of mechanism is at its clearest and most plausible when we consider Frankfurt-style examples. When Jones votes for the Democrat without intervention by Black, his action issues from "the normal faculty of practical reasoning",[12] which is appropriately reasons-responsive. But in the alternative scenario, in which Black does intervene, Jones's vote does *not* issue from his own faculty of reasoning. Fischer and Ravizza claim that Jones's vote issues from two different kinds of mechanism in the two scenarios.[13]

In less exotic cases, which don't involve Frankfurt-style neurosurgery, it's much less clear how we are to decide whether two mechanisms are the same, or of the same kind.

> We must confess that we do not have any general way of specifying when two kinds of mechanism are the same. This is a potential problem for our

[7] According to the family of arguments usually known as the 'Consequence Argument', agents *never* have alternate possibilities if determinism obtains. If that is correct, then regulative control is incompatible with determinism.

[8] Fischer and Ravizza 1998:241n2. I take it that this statement is oversimplified: it ignores *traced* moral responsibility, as I'll discuss in section 6.5. Guidance control *is* compatible with determinism.

[9] Fischer and Ravizza 1998:38. This phrasing is slightly open-ended, perhaps because Fischer and Ravizza intend their theory to be compatible with both physicalism and mind-body dualism. I will generally speak of mechanisms as though they are physical processes in the agent's brain. But I will not assume that physicalism is true, in any of the points I'll rely on in my arguments.

[10] Fischer and Ravizza 1998:241-2. To my knowledge no other examples of kinds of mechanism are given.

[11] Fischer and Ravizza 1998:86.

[12] Fischer and Ravizza 1998:38. It's not very clear from the text whether this phrase picks out a mechanism or a *kind* of mechanism; I assume it is the latter.

[13] In my view, Jones's vote in the counterfactual scenario is *not Jones's action* at all, since it is controlled by Black. However, I'll ignore that point here.

> approach; … rather than attempting to say much by way of giving an account of mechanism-individuation, we shall simply rely on the fact that people have intuitions about fairly clear cases of "same kind of mechanism" and "different kind of mechanism".[14]

The mechanism is central in the Fischer-Ravizza theory, because it figures in both of the requirements for guidance control. When the agent has guidance control over an action, the mechanism which actually operates to produce the action must be *reasons-responsive*, and the mechanism must be *the agent's own*.

6.3. OWNERSHIP

Guidance control requires "ownership of the mechanism that actually issues in the relevant behavior".[15] Ownership of an individual mechanism is attained via ownership of the relevant *kind* of mechanism. In turn,

> individuals *make* certain kinds of mechanism *their own* by *taking responsibility* for them. (When we speak of taking responsibility for a kind of mechanism, we understand this as "shorthand" for taking responsibility for behavior that issues from that kind of mechanism).[16]

For a person to take responsibility for a mechanism he must meet three further conditions. First, he "must see himself as the source of his behavior … [he] must see himself as an agent; he must see that his choices and actions are efficacious in the world".[17] Second, he "must accept that he is a fair target of the reactive attitudes as a result of how he exercises this agency".[18] Third, his "view of himself as an agent and sometimes appropriately subject to the reactive attitudes [must] be grounded in his *evidence* for those beliefs".[19]

Taking responsibility for mechanisms is part of the process of moral education and development which begins in childhood. Once achieved, guidance control can be exercised over actions which issue from the same kind of mechanism without ownership being constantly reviewed.[20]

> When one takes responsibility, at a certain point in one's life, for a certain kind of mechanism, this functions as a kind of "standing policy" with respect to that kind of mechanism. So, for example, if one has in the past taken responsibility for the mechanism of "ordinary practical reasoning"[21] (and in the absence of reconsideration of this mechanism), it follows that one takes responsibility for the currently operating mechanism of ordinary practical reasoning: taking responsibility is, as it were, *transferred* via the medium of "sameness of kind of mechanism".[22]

---

[14] Fischer and Ravizza 1998:40.
[15] Fischer and Ravizza 1998:241.
[16] Fischer and Ravizza 1998:241. I take it that the phrase "taking responsibility" is shorthand for "taking *moral* responsibility" – see Fischer and Ravizza 1998:184-201.
[17] Fischer and Ravizza 1998:210.
[18] Fischer and Ravizza 1998:211. Eshleman (2001) argues that this second condition is not appropriate, because (very roughly) an agent's beliefs about whether he is a fair target of reactive attitudes may be mistaken. I won't pursue this interesting issue, since it is somewhat tangential to the points I want to make.
[19] Fischer and Ravizza 1998:213.
[20] It is apparently possible for an agent to renounce ownership of a kind of mechanism, but (as far as I'm aware) no examples of this are given.
[21] This seems to be loose phrasing: "ordinary practical reasoning" here must be *a kind of mechanism*, rather than *a mechanism*.
[22] Fischer and Ravizza 1998:242.

The conditions of ownership are designed to help us distinguish causally determined actions from actions under manipulation.[23] In the Frankfurt-style examples involving Jones, the mechanism which would have issued in the action, if the neurosurgeon Black had intervened to manipulate him, is not a mechanism which Jones owns.[24] Therefore if Black were later to cause Jones to act in a certain way, using that mechanism, Jones cannot be morally responsible for that action. Determinism, however, does not prevent ownership of mechanisms. All of the conditions of ownership discussed above are compatible with determinism.

The ownership conditions also represent an attempt to give proper consideration to the agent's history, when assessing moral responsibility. This is an important advantage of the Fischer-Ravizza theory compared with Frankfurt's, which I criticised on this issue in section 2.3.2.[25]

6.4. REASONS-RESPONSIVENESS

Reasons-responsiveness is a dispositional property. Something is reasons-responsive in the actual situation if it *would* respond to reasons in counterfactual scenarios.[26]

In a Frankfurt-style case, *the agent* is *not* reasons-responsive in the Fischer-Ravizza sense. No matter what reasons he might have considered, Jones would still have voted for the Democrat candidate (because Black would have intervened to ensure this outcome). Jones could not have done otherwise.[27] According to Fischer and Ravizza, this shows that reasons-responsiveness of the agent himself is not necessary for moral responsibility (since they hold that agents in Frankfurt-style cases are morally responsible for their actions).

Instead, the agent has guidance control over his action if *the mechanism* which issues in the action is appropriately reasons-responsive (and he owns the mechanism). Fischer and Ravizza characterise three forms of reasons-responsiveness: strong, weak and moderate. The "moderate" form is the one they regard as necessary and sufficient for (direct) moral responsibility. This is best explained by first describing the other two forms.

They define the *strong* form as follows:

> Suppose that a certain kind *K* of mechanism actually issues in an action. Strong reasons-responsiveness obtains under the following conditions: if *K* were to operate and there were sufficient reason[28] to do otherwise, the agent would *recognize* the sufficient reason to do otherwise and thus *choose* to do otherwise and *do* otherwise.[29]

When testing for strong reasons-responsiveness, we must "ask *what would happen* if there were a sufficient reason to do otherwise".[30] Given the actual kind of mechanism used,

> the nonactual possible worlds that are germane to strong reasons-responsiveness are those in which the agent has a sufficient reason to do

[23] Stump (2002) objects that an agent could be manipulated but still meet the ownership conditions; see also Fischer and Ravizza's reply (2004:226-230). This issue does not affect my main argument.
[24] Fischer and Ravizza 1998:227-8.
[25] See also Fischer and Ravizza 1998:194-201.
[26] Fischer and Ravizza use "actual situation" interchangeably with "actual world", and "counterfactual scenario" interchangeably with "nonactual possible world". I take it that these terms are indeed equivalent; I will usually refer to situations and scenarios rather than "worlds".
[27] Fischer and Ravizza 1998:39.
[28] Fischer and Ravizza say that for the term "sufficient reason" we should read "justificatorily sufficient reason", and not "motivationally sufficient reason" (1998:41n13).
[29] Fischer and Ravizza 1998:41. They take no stand on debates about the ontological nature of reasons (1998:68n11).
[30] Fischer and Ravizza 1998:44 (my italics).

> otherwise … which are *most similar* to the actual world. (Perhaps there is just one such world, or perhaps there is a sphere of many such worlds.)[31]

Strong reasons-responsiveness is not necessary for moral responsibility. Suppose Jennifer decides to go to a basketball game tonight, and goes.[32] But if there had been a sufficient reason to stay at home instead – such as a deadline for a piece of work – she would have been "weak-willed", and she would still have gone to the game. Her actual-sequence mechanism is not strongly reasons-responsive, and yet we would hold her morally responsible for her action in the actual situation.[33]

For *weak* reasons-responsiveness,

> we (again) hold fixed the actual kind of mechanism, and we then simply require that there exist *some* possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise.[34]

Once again the relevant possible worlds are those in which an action issues from the same kind of mechanism as the actual-sequence mechanism. It's important to note that: "This possible world *need not* be the one (or ones) in which the agent has a sufficient reason to do otherwise … which is (or are) *most similar* to the actual world".[35]

Suppose Jennifer would not decide to go to the basketball game if the ticket were priced at $1000. Her actual-sequence mechanism is revealed to be at least weakly reasons-responsive by the fact that "there exists *some* scenario in which the actual mechanism operates, she has sufficient reason not to go to the game, and she doesn't go".[36]

Weak reasons-responsiveness (in the agent's actual-sequence mechanism) is necessary for (direct) moral responsibility. But it is *too* weak, by itself, to be sufficient: Fischer and Ravizza consider three respects in which it needs to be strengthened. First, in the possible scenarios being considered, the agent must do otherwise *because* there is sufficient reason to do otherwise (and not merely coincidentally with the existence of such a sufficient reason). Second, the agent must not merely respond in bizarre or incomprehensible ways to reasons in counterfactual scenarios; his mechanism must display an understandable "pattern" of reasons-responsiveness. (I'll discuss, below, an example which fails this requirement.) Third, the agent must respond to *moral* reasons and not merely to prudential reasons (this is one difference between adults and young children). Thus a *moderate* form of reasons-responsiveness is required, which incorporates these elements and lies somewhere between the strong and weak forms.

To specify what the moderate form requires, Fischer and Ravizza distinguish two constitutive elements of reasons-responsiveness: "receptivity to reasons" and "reactivity to reasons". Receptivity is a matter of recognising reasons. Reactivity requires translating reasons into both choices and behaviour. Moral responsibility, they argue, requires "only a very weak sort of reactivity*"* but "a *stronger* sort of receptivity" to reasons.[37]

---

[31] Fischer and Ravizza 1998:44n16.
[32] The example is from Fischer and Ravizza 1998:42-3.
[33] The "actual-sequence mechanism" is the one which issues in the action actually performed; "alternative-sequence" mechanisms would issue in actions in counterfactual scenarios.
[34] Fischer and Ravizza 1998:44.
[35] Fischer and Ravizza 1998:44n17 (my italics).
[36] Fischer and Ravizza 1998:45. I think there is a mistake in this quote: the phrase "the actual mechanism operates" should instead read "the same *kind of* mechanism operates". This same switch between talking about *mechanisms* and *kinds of mechanism* occurs several times in Fischer and Ravizza's discussion.
[37] Fischer and Ravizza 1998:69.

They use an example to explain and support this "asymmetry".[38] Brown takes a drug called 'Plezu', which causes euphoria but also lethargy and the likelihood of eventual loss of his job, family and self-respect: there is a sufficient reason to stop taking it.

Suppose first that Plezu is "nonaddictive": users do not face *irresistible* urges to take it.[39] But the only scenario in which Brown would stop taking Plezu is that in which he is told that doing so would cause his death. Here his mechanism is *weakly reactive* to reasons: there is at least one scenario in which he would refrain from taking the drug. Therefore he can be held morally responsible for taking the drug.

Now suppose that Brown faces an irresistible urge to take Plezu: there is no scenario in which he would refrain from taking it. Here, Brown's mechanism is not even weakly reactive, and so he cannot be morally responsible for taking the drug.

Finally, suppose that Brown would refrain from taking Plezu if his next fix cost $1000 (the urge to take it is not irresistible). But he would not refrain from taking the drug if the next fix cost $2000. Here his mechanism is not *receptive* to reasons in an understandable or appropriate "pattern", and therefore he cannot be held morally responsible for taking the drug. The mechanism must be "regularly receptive". This requires that, in counterfactual scenarios in which the same kind of mechanism operates,

> the agent would *recognize* reasons (some of which are moral) in such a way as to give rise to an understandable pattern (from the viewpoint of a third party who understands the agent's values and beliefs).[40]

In summary: a mechanism must exhibit both "weak reactivity" and "regular receptivity" in order to be moderately reasons-responsive. On the Fischer-Ravizza theory, ownership of a moderately reasons-responsive mechanism is necessary and sufficient for guidance control, and hence direct moral responsibility. From now on, following Fischer and Ravizza, I'll use the term "reasons-responsive" as a shorthand for "moderately reasons-responsive".

6.5. TRACING

Fischer and Ravizza do not use the term "direct" moral responsibility, nor any equivalent term. In fact they claim explicitly that guidance control is necessary and sufficient for moral responsibility for action.[41] But I think what they have in mind is revealed when they make a "refinement" to their account to incorporate "a tracing approach".[42]

According to Fischer and Ravizza, moral responsibility is traced from an earlier action at time *T1* to a later action at time *T2* when the following conditions are met:[43]

(a) "an agent's act at a time *T1* issues from a reasons-responsive sequence, and this act causes his act at *T2* to issue from a mechanism that is not reasons-responsive", and
(b) the agent "can reasonably be expected to have known" that the action at *T1* would (or might) "lead to" the action at *T2* issuing from a mechanism that is not reasons-responsive.

These conditions are very similar to the rough and open-ended condition I proposed in section 1.2.[44] I gave the example of Trisha who took a fix of a drug for the first time yesterday,

[38] Fischer and Ravizza 1998:69-70.
[39] As I'll discuss below in section 6.6.3, this is a very unusual definition of "nonaddictive".
[40] Fischer and Ravizza 1998:243-4. (This statement in the concluding chapter of the book is rather more precise than the "still quite vague" account discussed on pages 70-73.)
[41] I think this must be a misleading oversimplification – see footnote 8 above. They must mean that guidance control is necessary and sufficient for *direct* moral responsibility.
[42] See Fischer and Ravizza 1998:49-51.
[43] Fischer and Ravizza 1998:50.

knowing it to be extremely addictive. She now has a very strong craving the drug, and takes another fix this morning. Suppose that the mechanism which issues in this action is not reasons-responsive, and so Trisha does not exercise guidance control over this action. She is not directly morally responsible for it. But Trisha did exercise guidance control over her fix-taking action yesterday, and could reasonably be expected to have known that it might lead to her fix-taking today. Despite her lack of guidance control today, Trisha is morally responsible for injecting the fix *today* – via the tracing principle.

I take it from this discussion of tracing that, on the Fischer-Ravizza theory:

1. Guidance control of action *A* is necessary and sufficient for *direct* moral responsibility for action *A*.
2. If the agent does not have guidance control of action *A*, then, for (traced) moral responsibility for action *A*, it is necessary and sufficient that the agent had guidance control over some previous action *from which that moral responsibility for action A can be traced*.

I make no criticisms of Fischer and Ravizza's account of tracing.[45] The tracing principle seems to me extremely plausible. But, as I noted in section 1.2, I don't need to assume that it is true. I'll continue to distinguish direct from traced moral responsibility. In quotations from Fischer and Ravizza, I'll generally take it that "moral responsibility" means *direct* moral responsibility.

6.6. CRITICISMS OF THE FISCHER-RAVIZZA ACCOUNT

In section 6.7 I'll argue that actions with subverting causes are counterexamples to Fischer-Ravizza theory, and so guidance control is not sufficient for direct moral responsibility. Before then, I'll discuss a series of different criticisms which support that same conclusion. I'll argue that we do not have the intuitions which would support Fischer and Ravizza's claims about mechanism individuation (section 6.6.2). I'll then argue that the Fischer-Ravizza theory gives implausible treatments of both addictions (section 6.6.3) and compulsions (section 6.6.4). Finally I'll argue that their theory's definition of weak reactivity is simply *too weak*, since reactivity to certain reasons does not indicate moral responsibility (section 6.6.5).

But I'll begin by discussing a problem with their definition of weak reactivity which, as far as I'm aware, has not been pointed out by other commentators.

6.6.1. A problem with definitions

I think there's a flaw in the definition of moderate reasons-responsiveness, which is supposed to be necessary for guidance control. For weak reactivity:[46]

> [we] hold fixed the actual kind of mechanism, and we then simply require that there exist *some* possible scenario (or possible world) in which there is a sufficient reason to do otherwise, the agent recognizes this reason, and the agent does otherwise.[47]

---

[44] The condition I proposed was: "An agent bears traced moral responsibility for action A1 if: she bears direct moral responsibility for an earlier action (or omission) A0; action A1 results from A0; and the resultant occurrence of A1 was reasonably foreseeable by the agent, at the time of A0."

[45] They acknowledge that their account of tracing is "a sketchy and incomplete treatment of difficult issues" (1998:50).

[46] This condition of "weak reactivity", strengthened with the conditions for "regular receptivity", entails "moderate reasons-responsiveness". See section 6.4.

[47] Fischer and Ravizza 1998:44. (I quoted the same passage in section 6.4.)

Fischer and Ravizza specify only four elements which must obtain in the requisite possible scenario: a sufficient reason to act otherwise, recognition of that reason by the agent, the agent's acting otherwise (because of that reason), and the same kind of mechanism issuing in the action. I think there is a fifth requirement, which is not stated: the counterfactual scenario must contain *the most important reasons for which the agent performs her action in the actual situation.*

Here is an example which demonstrates the problem I have in mind. Ray is a man who is often angry, and today he is in a particularly angry mood. It seems as though everything has gone wrong since he got up this morning, and he's just had a blazing row on his mobile phone with his girlfriend. Just then a passing stranger, Nick, says to him: "Cheer up mate – it might never happen!". Ray punches Nick. Suppose that there is no scenario in which Nick says these mocking words and Ray does not punch him. Ray's anger is such that he would respond in this way in every counterfactual scenario, even if someone quickly offered him £10 million in the moment before he punched Nick. Ray's mechanism is not moderately reasons-responsive in the way required for guidance control.

Now consider a counterfactual scenario in which Nick says "good morning" to Ray as he walks past. Ray scowls at Nick but does not punch him. This is a possible scenario in which the four Fischer-Ravizza conditions are satisfied: a sufficient reason to act otherwise (Nick's pleasant greeting), recognition of that reason by the agent, the agent's acting otherwise (in response to that reason), and the same kind of mechanism issuing in the action. But this does not show that Ray's mechanism is moderately reasons-responsive *in the actual situation*. For, *given that Nick speaks with the actual words*, there is no counterfactual scenario in which Ray would act otherwise than punching Nick. My proposed fifth condition is required, I think, for moderate reasons-responsiveness.

Why don't Fischer and Ravizza include this fifth requirement? Perhaps they think it can be taken for granted, and doesn't need to be stated. For normal actions (where the agent has guidance control) there will always be *some* scenario in which the most important reasons for the actual action remain, and some additional reason prompts a change in response. In Fischer and Ravizza's example, Jennifer goes to the basketball game (presumably because she enjoys basketball and wants to see this particular game). Her decision would have been the same (although "weak-willed") if she had been given a deadline for a piece of work. But there will always be some counterfactual scenario in which she decides not to go to the game – in this case the price of a ticket rising to $1000 would suffice. I assume this is the reason why Fischer and Ravizza don't specify anything else about the counterfactual scenarios which are relevant.

I can see two objections to my argument. First, it might be objected that the mechanism in the counterfactual scenario is not of the same kind as in the actual situation. My reply is that I cannot see any reason to think that these mechanisms are of different kinds. In both scenarios Ray is in an angry mood before Nick appears, and he reacts accordingly.

The second objection might be that my example is unrealistic. One might think that there is bound to be *some* counterfactual scenario in which Ray would respond differently and not punch Nick – perhaps if he were distracted by a bolt of lightning, for example.[48] The problem with this second objection is this: if such counterfactual events as distraction by lightning are to be counted as establishing moderate reasons-responsiveness, it becomes very hard to see how any mechanism could *ever* be less than moderately reasons-responsive. This same theme recurs in the next two sections.

A problem with my suggested fifth condition is that it is rather imprecise – how should we define "the most important reasons for which the agent performs her action"? I don't

[48] Presumably a counterfactual scenario in which Ray were *hit* by the bolt of lightning and *prevented* from punching Nick would not be a relevant scenario, because the same kind of mechanism could not operate.

know how to solve this problem, but of course my main concern is simply to point out that it needs to be addressed – and isn't addressed by Fischer and Ravizza.

6.6.2. The theory relies on spurious intuitions about mechanisms

Michael McKenna criticises Fischer and Ravizza for providing "no principled basis for individuating mechanisms".[49] Instead, (as I discussed in section 6.2) they offer only an appeal to people's intuitions about whether mechanisms belong to the same kind. Two of his objections in particular indicate why McKenna finds this appeal unsatisfactory.

The first objection is that Fischer and Ravizza's intuitions about sameness of mechanisms simply reflect their own compatibilist convictions. There are many possible notions of sameness of mechanism across counterfactual scenarios. One of these would require that "the entire complex of proximal antecedent events and states" in the world before the action be held constant. Using this "stringent" notion of sameness, if determinism is true then the agent would perform the same action in *every* counterfactual scenario where the same mechanism issues in action.[50]

Fischer and Ravizza had attempted to head off this challenge, by claiming that the stringent view of sameness would only be appealing to someone with a "prior commitment to the view that causal determinism is incompatible with moral responsibility".[51] They seem to be implying that it is question-begging to take the stringent view. McKenna's charge is that it's equally question-begging for compatibilists to reject the stringent view.

McKenna's second objection is that our judgements about the sameness of mechanisms are likely to vary according to our explanatory perspective. A neurospychologist may be interested in physical states and processes in the agent's brain, when making her judgement; whereas in everyday discourse we don't refer to such states and processes, and we rarely consider their role at all.

Fischer replies to McKenna's concerns by conceding that his defence of mechanism individuation is "incomplete", and that "it is obvious that the notion of 'mechanism leading to action' is quite vague in itself".[52] But, for Fischer, a simple appeal to everyday intuitions – rather than general principles – need not be a weakness in a philosophical theory. His objective is to provide a theory which correctly reflects our intuitions about moral responsibility – the ones which we employ in everyday discourse. A principled account of mechanism individuation would be desirable; but Fischer's objective could be achieved without one. I think this is adequate as a defence against McKenna: Fischer's position here is structurally sound.

However, Fischer's position relies instead on two crucial appeals to our intuitions. First, it must be that the claims he makes about mechanism individuation do indeed match our intuitions. Second, the theory as a whole, incorporating its mechanism individuation component, must give verdicts about moral responsibility which match our intuitions.

In fact I think there are fatal problems in both areas. Although it succeeds in giving verdicts on moral responsibility which match our intuitions about most actions, there are nevertheless important counterexamples to the Fischer-Ravizza theory which I'll consider in the next two sections. Furthermore, the intuitions about mechanism individuation which are required to support the theory are largely spurious. Most importantly, I think the following claim made by Fischer and Ravizza is false:

[49] McKenna 2001:96.
[50] McKenna 2001:96-97.
[51] Fischer and Ravizza 1998:52n22.
[52] Fischer 2004:240.

> people have intuitions about fairly clear cases of 'same kind of mechanism' and 'different kind of mechanism'.[53]

Evidence in support of my view comes from Fischer's exchange with Eleonore Stump over the conditions for ownership of mechanisms. Stump objects that the ownership conditions in the Fischer-Ravizza theory are inadequate – it could be that they are satisfied while the agent is manipulated and so not morally responsible for his action.[54] I'm not concerned here with the details of this objection. More relevant for my argument is that Fischer, in his reply to Stump, says that she "employs an overly broad notion of mechanism individuation" in her critique.[55] Further, he thinks

> it should be evident that, in order to render the Fischer-Ravizza account of manipulation cases even minimally plausible, we are not thinking of the relevant mechanisms as individuated so broadly [as Stump does].[56]

He seems genuinely surprised that Stump could have misinterpreted the theory on "this very basic point".

My point is this. Whatever the merits of their respective arguments about ownership, it's clear that Stump and Fischer have formed very different views about *mechanism individuation*. And this is despite the detailed study of the Fischer-Ravizza account which Stump clearly must have made before proposing her critique. This already is evidence that we do not have reliable intuitions about how to compare kinds of mechanism. Nor can we set aside Stump's apparent misunderstanding as an aberration, since Harry Frankfurt endorsed her critique of Fischer.[57]

Speaking for myself, I don't think I ever have any intuitions about whether two actions issue from the same or different kinds of mechanism. This is so whether the actions are mine or someone else's. In fact, I don't think I have any intuitions at all about mechanisms, nor about "kinds" of mechanism – even while reading the Fischer-Ravizza example cases which are supposed to draw out such intuitions.

The role of the very term "mechanism" is curious here. If a mechanism is nothing over and above a process, why does the theory not draw on our intuitions about "process individuation"? I think the answer must be that it's even less plausible to suppose that we have intuitions about how to compare processes and kinds of process.

A Frankfurt-style case is the only one in which I feel an intuition that two different kinds of *process* are being compared. Here I do have an intuition that a process involving a microchip is a different kind of process from one involving only organic circuitry in the agent's brain. But this intuition alone falls far short of validating the claims about mechanism individuation in the Fischer-Ravizza theory.

A further problem concerns this claim made by Fischer and Ravizza:

> Imagine that the agent somehow gets considerably more energy or focus if he is presented with a *strong* reason to do otherwise, and it is only in virtue of these factors that he successfully reacts to the reason. There certainly can be cases like this, but it is natural to say that, when the agent acquires significantly more "energy or focus," this gives rise to a *different mechanism* from the actual mechanism.[58]

[53] Fischer and Ravizza 1998:40. (I used the same quote in section 6.2.)
[54] Stump 2002:46-56.
[55] Fischer 2004:230.
[56] Fischer 2004:229.
[57] Frankfurt 2002a:61. (This was published before Fischer's response to Stump.)
[58] Fischer and Ravizza 1998:74.

From the context it's clear that they think a different *kind of* mechanism would arise, not merely "a different mechanism".[59] I think this claim is false. It does not seem natural to me to say that an agent's mechanism is of a different kind when he gains significantly more focus. Insofar as I can gain any intuitive grasp on the notion of kinds of mechanism, it seems to me that the same kind of mechanism is involved in both cases.

McKenna wonders why it isn't *the mechanism,* rather than the agent, which can gain more energy or focus when there is a strong reason to do otherwise.[60] This may seem an odd way to speak; but speaking of mechanisms *at all* is already very odd. My intuitions don't favour either way of speaking over the other. Insofar as I can grasp the notion of mechanisms at all, I would say that *an agent* could gain more energy or focus, and yet still act from the same kind of mechanism.

In summary, the notions of mechanisms and kinds of mechanisms are nebulous and unclear; we do not have the intuitions which would support Fischer and Ravizza's claims about mechanism individuation. The authors don't provide any principles or other means to individuate mechanisms, and so their claims about sameness of kinds of mechanism are effectively unsupported. Since those claims are central to the definition of guidance control, this is a significant flaw in their theory.

6.6.3. The theory gives an implausible treatment of addictions

Meanwhile, the theory deals poorly with actions performed by addicts. In section 6.4 I described the example of Brown, who takes a drug called 'Plezu' which is likely to lead to the loss of his job and family. Fischer and Ravizza describe Plezu as "nonaddictive", because Brown's urge to take it is not *irresistible*. In fact, the only scenario in which Brown would stop taking Plezu is that in which he is told that doing so would cause his death.

I take it that the Fischer-Ravizza usage of the term "nonaddictive", and hence too of the term "addictive", is highly unusual. It must be doubtful that addiction *ever* exists, on this definition.[61] As far as I'm aware, Fischer and Ravizza offer no evidence that any addictions or compulsions are irresistible in the sense that even the weak reactivity condition cannot be met.[62]

Fischer and Ravizza claim that Brown is morally responsible for his action. Suppose Brown objects as follows:

> "whereas I would have responded to a very different incentive for doing otherwise, the mechanism on which I acted did not – and *could not* – have responded to the *actual* incentive to do otherwise. Given this, it is unfair to hold me morally responsible."[63]

Fischer and Ravizza respond to Brown's objection by claiming that "reactivity is all of a piece". If a mechanism of the same kind would react differently to some incentive to do otherwise, this shows (they think) that the actual-sequence mechanism "*can* react to *any* incentive to do otherwise".[64] Specifically they say:

---

[59] The authors often use the term "mechanism" where they must mean "kind of mechanism". It may be that they do so intentionally as a form of shorthand (though to my knowledge they do not explain such a move). A potential danger is that an appeal might be made to the wrong set of intuitions. For example, intuitions about mechanisms (if indeed we have any) may not support conclusions about *kinds* of mechanism.

[60] McKenna 2001:99.

[61] Watson (2001) makes this point.

[62] Brown's mechanism is weakly reactive because there is at least one counterfactual scenario in which he would refrain from taking the drug (via the same kind of mechanism).

[63] Fischer and Ravizza 1998:73 – these words are spoken by the fictional Brown.

[64] Fischer and Ravizza 1998:73.

> it is plausible to reply to Brown that his mechanism of practical reasoning (the mechanism that actually produced his behavior) could in fact have reacted to his *actual* reason not to take Plezu.[65]

I think there are two convincing objections to this claim. The first is simply a denial of the quotation just given: it is *not* plausible to conclude, from the fact that a mechanism of the same kind would react differently to a different reason in a counterfactual scenario, that the *actual* mechanism can react differently. This is a complex issue, and somewhat tangential to my main argument, so I mention it only briefly and won't rely on this objection.[66]

The second objection fits well with my main argument: the "weak reactivity" requirement which forms part of the guidance control condition is much *too* weak. That there is a single counterfactual scenario in which an agent would act otherwise (via the same kind of mechanism) is not enough to establish that he is morally responsible for his action.

Consider an even more extreme example of addiction. Harry is a long-term heroin user. He takes heroin this morning, and he would take it even in a counterfactual scenario in which doing so would cause his own death.[67] The *only* counterfactual scenario in which he would not take the drug, is one in which a kidnapper would torture and then murder all of his children. Harry is clearly very strongly addicted to heroin, in any normal sense of the term "addicted". It's not plausible to hold Harry directly morally responsible for taking the drug, even though his mechanism meets the weak reactivity condition.

One might object to that claim by pointing to a contrary intuition: that heroin addicts do bear moral responsibility for their actions. I recognise this intuition and feel its pull to some extent. Much of its strength, though, is attributable to the tracing principle. Perhaps Harry was DMR when he tried the drug for the first time, and the resultant occurrence of his addicted actions was reasonably foreseeable by him. It's very plausible, then, to hold Harry morally responsible for his taking the drug this morning, via tracing. But what is not plausible, I think, is to say that Harry is *directly* morally responsible for this action.[68]

A separate problem is that it's far from clear that the mechanism which issues in Brown's action is, as Fischer and Ravizza claim in the quotation above, "his mechanism of practical reasoning". Setting aside the point that "practical reasoning" is surely a *kind* of mechanism rather than a particular mechanism, this conflicts with my intuitions about actions under severe addiction. If Brown (or Harry) craves his drug so strongly that there is only one reason which would prevent him taking a fix, then it seems much more plausible that the operative mechanism is of a very different kind from those which issue in their other everyday actions. In fact, I find it plausible to imagine that Brown (or Harry) might well not meet the *ownership* conditions for the kind of mechanism which operates when he takes a fix to relieve a craving. That is, he may well not have "taken responsibility", in the appropriate way, for previous actions issuing from the same kind of mechanism. There is thus a plausible means, available to Fischer and Ravizza, by which to judge that Brown and Harry are not morally responsible for their actions. I find it strange that they do not do so, and that instead they reach a counterintuitive verdict. This reinforces my conviction that their claims about sameness of mechanisms are not supported by our intuitions.

To conclude, it's not plausible to hold Harry or Brown directly morally responsible for taking their drugs. Since both apparently meet the conditions of guidance control, I conclude that

[65] Fischer and Ravizza 1998:74 (my italics).
[66] For more details see Watson 2001:299-300, and Fischer and Ravizza 2004:242-244.
[67] He recognises that this as a sufficient reason not to take the drug (he does not want to die).
[68] This is one of many instances, I think, in which our intuitions do not distinguish very clearly between traced moral responsibility and direct moral responsibility. We must take care when judging which one is applicable.

guidance control is not sufficient for direct moral responsibility.[69] In particular, the reactivity part of the reasons-responsiveness requirement is *too weak*. That a mechanism could react differently in a single counterfactual scenario does not establish that the agent is morally responsible, even if all the other conditions of guidance control are met.

6.6.4. The theory gives an implausible treatment of compulsions

A similar argument can be made about the theory's treatment of agents with compulsions. Al Mele raises the example of Fred, who is agoraphobic and hasn't left his house in ten years.[70] He stays at home rather than attend his beloved daughter's wedding in the church next door. But if Fred's house had caught fire on that afternoon, he would have been even more terrified of the flames, and would have left the house and gone to the church. Fred's mechanism, which issues in his actual action of watching TV at home during the wedding, is weakly reasons-*reactive*.[71] Thus (provided that it is regularly reasons-receptive and owned by Fred) the mechanism is moderately reasons-responsive; and Fred is morally responsible for his actual action on the Fischer-Ravizza account. But this verdict is profoundly counter-intuitive, as Mele notes. If Fred's agoraphobia is so extremely debilitating that it would take a raging fire to move him to leave the house, then it's much more plausible to excuse him from moral responsibility.

Fischer and Ravizza address this point directly in a paper published in the same symposium as Mele's.[72] One might have expected them to claim that Fred's mechanism in the fiery counterfactual scenario is of a different kind than the actual-sequence mechanism. In section 6.6.2 I discussed their claim that, if an agent would gain more "energy or focus" when presented with a "strong" reason to do otherwise, then a different kind of mechanism would operate. I rejected that claim as unsupported by intuitions. Nevertheless I find it surprising that they don't take that line in the case of Fred, which seems closely analogous. In fact, this adds to my conviction that their intuitions about sameness of mechanism are spurious.

Instead, Fischer and Ravizza take the following line. Fred is exempted from moral responsibility only if his phobia "issues in genuinely irresistible urges" – which in this case it does not.[73] The phobia is not "genuinely irresistible", on the Fischer-Ravizza account, if Fred's mechanism is weakly reasons-reactive. Fred can thus be held morally responsible; but it does not follow from this that he is blameworthy for his action. Indeed, Fischer and Ravizza themselves would not be inclined to hold him blameworthy, given Mele's description of the case.

On what grounds do Fischer and Ravizza defend their claim that Fred is morally responsible? They see Fred as "an *apt target* for the reactive attitudes on the basis of [his action]".[74] They draw a distinction between Fred on the one hand and, on the other, an agent who has irresistible urges or an agent who is "being significantly manipulated". I don't think that *being an apt target for the reactive attitudes* captures the distinction between Fred and those agents. The example with which Fischer and Ravizza first introduced the notion of

[69] Brown does meet the guidance control conditions, according to Fischer and Ravizza themselves; I'm assuming that they would make the same judgement on Harry.

[70] Mele 2000:450. I've adapted the example very slightly to make it a case about responsibility for an action rather than for an omission.

[71] The mechanism is weakly reasons-reactive if there is a single counterfactual scenario in which the same kind of mechanism issues in a different action in response to a sufficient reason.

[72] Fischer and Ravizza 2000b:470-472.

[73] Fischer and Ravizza 2000b:471. It's interesting that they describe the phobia as issuing in *urges*, rather than its issuing in actions. It's not clear to me whether they consider that a phobia *is* a mechanism (or a kind of mechanism), or whether they have some other relation in mind between phobias and mechanisms. Once again, I find I have no intuitions on the subject.

[74] Fischer and Ravizza 2000b:471.

being an apt candidate for reactive attitudes was taken from Wallace: a charming colleague who has cheated and lied can be an apt target of reactive attitudes, even though one may not actually feel resentment about his actions.[75] Here it is very clear that the colleague is an apt target of reactive attitudes; an acquaintance who has not yet been charmed is very likely to experience indignation or resentment about his actions. But that is not true in Fred's case. No amount of distance from Fred's case would make me regard Fred as an apt target of reactive attitudes. In fact, Fred's situation looks rather like the inverse of Wallace's case. I think it's possible that Fred's daughter might feel some indignation or other reactive attitudes toward him, while at the same time knowing that these emotions are misplaced – i.e. she might feel the attitudes while knowing that Fred is not an apt target for them. I don't find it plausible to say that Fred is an apt target of reactive attitudes, and so morally responsible for his action.

To conclude, I think it's flatly implausible to hold Fred morally responsible in this case. Since Fred does meet the conditions of guidance control, according to Fischer and Ravizza themselves, guidance control is not sufficient for (direct) moral responsibility. In particular, the reactivity part of the reasons-responsiveness requirement is *too weak*. That a mechanism could react differently in a single counterfactual scenario does not establish that the agent is morally responsible, even if all the other conditions of guidance control are met.

Furthermore, just as in the cases of the addicts Brown and Harry, it's far from clear that Fred acts from a mechanism of which he has taken ownership. Fischer and Ravizza seem to believe that Fred's action issues from "the mechanism of practical reasoning".[76] It seems more plausible to think that Fred's current action issues from a mechanism (or kind of mechanism) which has in the past issued in agoraphobic actions. But if that is the case, then it seems very unlikely that Fred would have "made this mechanism his own", in Fischer-Ravizza terminology, by "taking responsibility" for those past actions.

6.6.5. The sufficient reason to do otherwise may be another compulsion

Watson raises a related issue.[77] Suppose there is only one counterfactual scenario in which a heroin addict (I'll call him Adrian) would not perform his actual action of taking a fix of the drug. That counterfactual scenario is one in which Adrian's supply of heroin is put into a cage full of rats, of which he is profoundly phobic. Adrian's actual-sequence mechanism is moderately reasons-responsive, and so on the Fischer-Ravizza account (provided that that he owns the mechanism) Adrian has guidance control and is morally responsible for his action.

But here, as Watson points out, susceptibility to counterincentives is not evidence of control or moral responsibility. Rather, Adrian is "in the grip of competing compulsions". This kind of case is a counterexample to the Fischer-Ravizza guidance control conditions. The theory's definition of weak reactivity is simply *too weak* – it can be met when an agent clearly is not morally responsible for his action.

One objection to this argument, open to Fischer and Ravizza, is to claim that the mechanism in the "rat cage" counterfactual scenario is of a different kind from his actual-sequence mechanism, because Adrian would gain some "extra focus or energy" if faced by his phobia. In section 6.6.2 I argued that this claim is not supported by our intuitions.

A similar objection would be to claim that the two mechanisms are of different kinds because one involves an addiction and one involves a phobia. That claim seems to me very plausible. But making this objection would open another line of attack against the theory's

[75] See section 1.1.3.
[76] They do not say this explicitly in their reply to Mele, but this line would be consistent with what they say about Brown the Plezu addict – see section 6.6.3.
[77] Watson 2001:295.

use of "kinds of mechanism".[78] For Fischer and Ravizza have claimed, in several crucial cases which don't seem markedly different from Adrian's, that the same kind of mechanism does operate in relevant counterfactual scenarios.

For example, let's compare Adrian's case with two others. First, Fischer and Ravizza claim that Fred (the agoraphobic) has guidance control if he would react to a raging house fire. On what grounds can we confidently say that Fred's agoraphobic mechanism is of the same kind as his counterfactual fire-fearing mechanism? I can't see that the difference in kinds of mechanism between actual and counterfactual scenario is less significant in Fred's case than in Adrian's. A similar question arises over Brown's ability to stop taking Plezu in the counterfactual scenario in which he is told that he will otherwise die. Why believe that his Plezu-devoted mechanism is of the same kind as his death-fearing mechanism? I see no good grounds to believe that Brown's and Fred's mechanisms are of the same kind, while at the same time denying that Adrian's mechanisms are of the same kind. So this objection cannot help against my conclusion that Adrian's case is a counterexample to the Fischer-Ravizza guidance control conditions.

6.6.6. Conclusions from these criticisms

I draw three key conclusions from the criticisms I've discussed. First, the notions of "mechanisms" and "kinds of mechanism" are unclear, and crucial claims made about them by Fischer and Ravizza are not supported by intuitions. Second, the "weak reactivity" requirement, which forms part of moderate reasons-responsiveness, is simply *too weak*.[79] Third, there is a flaw in the Fischer-Ravizza notion of "ownership" of a kind of mechanism.

These conclusions are independent, in the sense that each one is well supported even if the others are wrong. For example, even if there are no flaws in the Fischer-Ravizza accounts of mechanisms and ownership, it remains the case that the weak reactivity requirement is too weak; and so on. All three of the problems are evident in most of the cases I've discussed, so it's difficult to disentangle them.

I'll summarise my case on each problem in turn, beginning with *mechanisms*.[80] In section 6.6.2 I discussed some of the claims about sameness of mechanisms which are most important for the Fischer-Ravizza theory, and argued that – contrary to the authors' assertions – they are not supported by our intuitions. Then in the next two sections (6.6.3 and 6.6.4) I showed that those claims don't help Fischer and Ravizza to deal with actions arising from addictions and compulsions.

My case against the Fischer-Ravizza *reactivity condition* is as follows.[81] To demonstrate weak reactivity, a mechanism need only react differently to any sufficient reason, in at least one counterfactual scenario. That reason may be very extreme, such as the threat of torture and murder of the agent's children. It could be even more extreme, such as the threat of the end of the world: for weak reactivity, there is no constraint on the kind of reason or counterfactual scenario to be contemplated. It simply is not plausible to say that this degree of reactivity is appropriate for judging moral responsibility. If some form of reasons-

[78] Watson makes this point (2001:297).

[79] A mechanism is weakly reasons-reactive if there is a single counterfactual scenario in which the same kind of mechanism issues in a different action in response to a sufficient reason. I have nothing to say about the "regular receptivity" element of moderate reasons-responsiveness, which is relatively uncontroversial.

[80] The central role played by mechanisms in their theory enables Fischer and Ravizza to give accounts of moral responsibility in Frankfurt-style examples, and in cases of manipulation of agents. I think both of those issues can be addressed without appeals to mechanisms, but I haven't pursued that point.

[81] Here I'm ignoring the definitional problem discussed in section 6.6.1.

responsiveness is indeed required for moral responsibility, the reactivity condition must be revised.[82]

Finally, there are problems with mechanism *ownership*. One problem, of course, is that if the notion of mechanisms is fatally unclear then so must be the notion of owning a mechanism. But there remains another problem, even if we assume that the notion of mechanisms can be made clear. I discussed examples of actions arising from addictions and compulsions, in which Fischer and Ravizza claim that the operative mechanism is owned (and the other conditions for guidance control are also met). I argued that the agents in these actions are *not* morally responsible, so that the actions are straightforwardly counterexamples to the claim that guidance control is sufficient for direct moral responsibility.

However, there is another way to interpret these cases, and defend a Fischer-Ravizza-style account of mechanism ownership, which I think is more plausible. We might instead say, contra Fischer and Ravizza's application of their own theory, that the agents in these actions *do not own* the operative mechanism. This could be the case if the mechanism issuing in those actions is not of the kind "ordinary practical reasoning". Then it may well be that the agent has not "taken responsibility" for the mechanism, because he has not taken responsibility for actions which have issued from mechanisms of the same kind in the past. To take this view seems more charitable to the Fischer-Ravizza theory than are the authors themselves – at least on this particular issue. On this view, the account of ownership of mechanisms is not undermined by cases of addiction and compulsion.

In the next section, though, I'll discuss another kind of case which definitely does undermine the Fischer-Ravizza account of mechanism ownership: actions with subverting causes. For these actions there is no doubt that the mechanism is owned, and the other conditions of guidance control are met, and yet the agent is not directly morally responsible.

6.7. ACTIONS WITH SUBVERTING CAUSES AND GUIDANCE CONTROL

For the remainder of the chapter I'll set aside the various criticisms I've already made of the Fischer-Ravizza theory of moral responsibility. I'll build an independent argument, based on actions with subverting causes, to show that the guidance control conditions are not sufficient for direct moral responsibility.

As a brief reminder, the five key cases of actions with subverting causes on which I build my main arguments were the following.[83] Samantha walks past a man slumped in doorway after being told that she is running late for her appointment to give a presentation. Martha, like the person sitting next to her, continues to fill out a market research questionnaire despite loud cries from the researcher in the next room. Judy judges that a candidate possesses a high degree of flexibility in solving problems, after being told that she will meet the candidate later. Ingrid interrupts a conversation shortly after completing a puzzle in which some of the words were connected with a theme of rudeness. Gina gives up trying to solve a difficult problem, after earlier resisting the temptation to eat some chocolate cookies.

Using these five examples of actions with subverting causes, I aim to demonstrate that the Fischer-Ravizza guidance control conditions are not sufficient for direct moral responsibility. In each of the five cases, I claim that the agent does exercise guidance control, but that she is not directly morally responsible for her action. In the next section, I'll defend

[82] I think it's very difficult to specify this reactivity condition. I have tried to formulate one, and rejected several candidates. I've come to suspect that no specification can be given which is genuinely independent of judgements about moral responsibility for actions. But that's not a claim I can defend here, so I remain uncommitted about whether some condition of reasons-responsiveness is indeed necessary for direct moral responsibility.

[83] See section 3.2 for more details.

my claim that the guidance control conditions are met in each of the five key cases. Then, in section 6.7.2, I'll set out my argument that the agents are not directly morally responsible for their actions.

It's important to note that the conclusion is true if the premises are true for *any* action. To refute the argument, an objector would need to show, for *all five* key cases, that one of the premises is false. A single counterexample will be enough to establish my argument.

My claim that subverting causes do occur is supported by the evidence I gave in section 3.2, introducing the five key cases. I also discussed a much wider range of experimental evidence for similarly subverting causes of actions (in section 3.6), and objections to my claim that there are actions with subverting causes (section 3.7). Furthermore, even if there are never in fact any subverting causes of actions, my argument nevertheless reveals something important about the conditions for direct moral responsibility. Simply considering the potential threat posed by subverting causes reveals that guidance control is not sufficient for direct moral responsibility. Guidance control might in practice be perfectly correlated with direct moral responsibility, but not logically sufficient for DMR.

6.7.1. Guidance control in the five key cases

A premise of my argument is that the agents in the five key cases do meet the Fischer-Ravizza guidance control conditions. There are several ways to object that they do not. Each objection would deny that an element of the guidance control conditions is present in one or more of the five actions. I'll consider each element in turn.

Someone might object that the agent's mechanism in the example actions is *not moderately reasons-responsive*. To be moderately reasons-responsive, a mechanism must display "weak reactivity" and "regular receptivity" to reasons.[84] In order to be weakly reactive, it must be the case that the same kind of mechanism would issue in a different action in at least one counterfactual scenario in which there is a different reason to act. There are two ways in which that might be false. First, it might be that there is *no counterfactual scenario* whatsoever – irrespective of the kinds of mechanism involved – in which a different action would occur. That is clearly not the case for any of the example actions.

The second way in which weak reactivity might be lacking is that different actions would only occur in counterfactual scenarios which involve *different kinds of mechanism* from the mechanism which issues in the actual action (Fischer and Ravizza call this "the actual-sequence mechanism"). It's clear that, in each of the example actions, there is a counterfactual scenario in which a different action issues from a mechanism that fits the Fischer-Ravizza description "ordinary practical reasoning".[85] The key question, then, is whether the actual-sequence mechanism is of that same kind. As far as I can see, there's no reason to believe that it's not. Only the fact that we happen to know of the influence of the subverting causes, in these particular cases, could make us think otherwise. If we did not know of the influence of the subverting causes – as the agents do not – then we would happily describe the actions as issuing from ordinary practical reasoning. In all likelihood, as I argued in chapter 3, there are many everyday actions which are influenced in similar ways; we would describe those actions as issuing from ordinary practical reasoning, too. So I think that the actual-sequence mechanism is weakly reactive to reasons, in all five example actions.

The "regular receptivity" condition requires that the agent's mechanism must display an understandable "pattern" of reasons- responsiveness. It's not enough for the agent to respond to reasons in bizarre or incomprehensible ways. The objector might claim that that the agents in my five examples are responding to reasons in bizarre ways. I deny this claim. I think that

[84] See section 6.4.

[85] For example, there are counterfactual scenarios in which the slumped man would call out to ask for help, and Samantha would offer him help via a mechanism of the kind described as "ordinary practical reasoning".

it's quite easy to understand, in all five cases, how the agents are responding to the reasons they encounter. The problem is not that their responses are bizarre; in fact they are all too understandable. The problem is rather that their actual responses are not the responses they would wish to make, if they learned of the influence of the subverting causes. Furthermore, the agents are all normal adults, whose mechanisms of ordinary practical reasoning respond in normal and unremarkable ways to everyday reasons. I see no grounds for doubting that the agents in my examples meet the Fischer-Ravizza receptivity requirements. I conclude that the agents' mechanisms in the example actions are indeed moderately reasons-responsive.

The other element of guidance control which might be missing, according to an objector, is *ownership* of the actual-sequence mechanism. Ownership of a mechanism is achieved by taking responsibility for actions which issue from a mechanism of the same kind.[86] This in turn involves a further condition: the agent must accept that she is a fair target of the reactive attitudes when she acts from that kind of mechanism.[87]

How might the objector deny that the agents in the five example actions own their mechanisms? He might begin by denying that the agents have taken responsibility for previous actions issuing from the same kind of mechanism. This is implausible, for the same reasons I discussed above: the operative mechanisms in every case fit the description "ordinary practical reasoning", and the agents are all normal adults who have taken responsibility for previous actions issuing from that kind of mechanism.

Instead, the objector might move on to deny that the agents *would take responsibility for the five particular actions* which I'm using in my argument – the actions which have subverting causes. I have two replies to this objection. The first is simply that responsibility for a mechanism is supposed (on the Fischer-Ravizza account) to be taken gradually over a period spanning many historical actions issuing from that kind of mechanism, and is not something which the agent takes or renounces for individual actions. By the time of their example actions the five agents are adults who have long ago taken responsibility for mechanisms of the kind "ordinary practical reasoning".

The second reply is that the agents of the five example actions *do* see themselves as morally responsible – and hence fair targets of reactive attitudes – for those actions. They do not know that their actions have subverting causes. The actions are just like many of our everyday actions, for which we all feel ourselves morally responsible. Without specialist knowledge, the agents do not even suspect that there may have been subverting causes. Thus I think both forms of this objection fail, and it's not plausible to claim that the agents do not own their actual-sequence mechanisms.

6.7.2. Subverting causes and direct moral responsibility

I've argued that the agents in the five key cases do meet the Fischer-Ravizza guidance control conditions. I'll now defend my claim that these agents are not directly morally responsible for their actions with subverting causes. In section 3.4.2 I argued for the same claim by drawing on key elements of Harry Frankfurt's position. There, I made the case that the effective desire in each of the five actions is not truly the agent's own, because she would not F-accept it if she knew all the causes of the action. If the effective desire is not truly the agent's own then she is not DMR for it, according to this identificationist axiom which Frankfurt holds:

> (4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).

[86] See section 6.3.

[87] Strictly, there are two further conditions of ownership, but these are true of all normal adult agents: she must see herself as an agent, and this belief must be appropriately grounded in evidence.

I concluded that followers of Frankfurt should accept that the agents in the five cases are not DMR for their actions.

Fischer and Ravizza, though, do not hold (4). So followers of Fischer and Ravizza will not be convinced by the argument that I made in section 3.4.2. Instead, I'll argue against the Fischer-Ravizza position directly, in a way which I think should persuade its advocates that the agents in the five cases are not DMR for their actions.

My argument focuses on the relationship between moral responsibility and ownership of a kind of mechanism in Fischer and Ravizza's theory. This relationship is best summarised in a pair of quotations:[88]

> individuals *make* certain kinds of mechanism *their own* by *taking responsibility* for them. (When we speak of taking responsibility for a kind of mechanism, we understand this as "shorthand" for taking responsibility for behavior that issues from that kind of mechanism).[89]

Since guidance control is sufficient for moral responsibility, to take responsibility for a kind of mechanism is to take responsibility for *all* the actions which issue from mechanisms of that kind when the other guidance control conditions are met.

> When one takes responsibility, at a certain point in one's life, for a certain kind of mechanism, this functions as a kind of "standing policy" with respect to that kind of mechanism. So, for example, if one has in the past taken responsibility for the mechanism of "ordinary practical reasoning" (and in the absence of reconsideration of this mechanism), it follows that one takes responsibility for the currently operating mechanism of ordinary practical reasoning: taking responsibility is, as it were, *transferred* via the medium of "sameness of kind of mechanism".[90]

(Strictly, what one takes responsibility for and sometimes later reconsiders is not a mechanism but a *kind of mechanism*.[91] To keep this point clear without unduly cumbersome phrasing, I'll refer to a kind of mechanism as "a mechanism-kind".) The authors do not clarify what "reconsideration" of a mechanism-kind involves or implies. But I assume they mean that an agent can *reject responsibility* for a mechanism-kind for which she has previously *taken responsibility*. In doing so (I take it) the agent would *reject ownership* of the mechanism-kind which she had previously *made her own* by taking responsibility for it.

Against Fischer and Ravizza, I've argued that we don't take responsibility for mechanisms or mechanism-kinds: we simply don't have intuitions about how to individuate these notions, which we would need in order to take responsibility for them.[92] I also argued that the use of mechanisms in the Fischer-Ravizza theory generates implausible verdicts about moral responsibility for actions involving compulsions and addictions.[93]

Nevertheless, the idea that agents *take responsibility* for certain actions does seem to reflect an important element of our practice of responsibility attribution. I'm inclined to think

[88] I used these same quotations in section 6.3 above.
[89] Fischer and Ravizza 1998:241. By "behavior" they mean to include both actions and omissions; I am focusing mainly on actions.
[90] Fischer and Ravizza 1998:242.
[91] The phrasing in the Fischer-Ravizza quote above is loose: "ordinary practical reasoning" is *a kind of mechanism*, rather than *a mechanism*. (I made the same point in footnote 21 above.)
[92] See section 6.6.2.
[93] See sections 6.6.3, 6.6.4, and 6.6.5.

that a more accurate account of taking responsibility would not feature *mechanisms* at all.[94] Instead it seems to me more accurate to say that we take responsibility for certain *kinds of action*; we do *not* do so indirectly, via taking responsibility for mechanism-kinds. But mechanisms are crucial to the Fischer-Ravizza theory, in two ways. In certain cases of manipulation of the agent, the theory states that an unowned mechanism issues in the action.[95] And it is reasons-responsiveness of *the mechanism*, rather than the agent, which enables moral responsibility for actions in Frankfurt-style cases. So I'll assume that the Fischer-Ravizza theory is correct, including its claims about mechanisms.

Even though, as I've argued, the notion of mechanisms has little support from our intuitions, the role they supposedly play in most everyday actions (which don't involve manipulation, addiction, compulsion and so forth) is relatively innocuous. The idea that we take responsibility for the mechanism-kind "ordinary practical reasoning" doesn't clash violently with our intuitions, because the kinds of actions we might expect to issue from such a roughly delineated mechanism-kind are also roughly the kinds of action for which we expect to be held responsible.

So I'll assume, for the sake of argument, that the agents in my five key cases have indeed taken responsibility for their mechanism-kind "ordinary practical reasoning", and hence for all the actions which issue from "ordinary practical reasoning".[96] But now let's consider what would happen if those agents knew all the causes of their actions in the five key cases. I think each agent would *reject* responsibility for her action. She would not accept that she is a fair target of reactive attitudes in relation to the action, if she learned of its subverting causes.

On Fischer and Ravizza's view, an agent's rejecting responsibility for actions which issue from a mechanism-kind presumably entails her rejecting responsibility for, and *ownership* of, that mechanism-kind. (This mirrors their position on *taking* responsibility and ownership of mechanism-kinds. But as far as I can tell they don't discuss the implications of *rejecting* responsibility for actions and mechanism-kinds.) In turn, rejecting ownership of a mechanism-kind must entail rejecting responsibility for all the actions which issue from that mechanism-kind, since ownership of the actual-sequence mechanism is required for guidance control and DMR. It would be foolish to claim that the agents in the five key cases would reject responsibility for all the actions which issue from their ordinary practical reasoning, just because they learned about the subverting causes of some of their actions. I do not claim that. Instead, what this discussion shows is that *taking or rejecting ownership of a mechanism-kind is much too inflexible* as a means of determining responsibility for individual actions.

Ownership of a mechanism-kind is much too far removed from the causes of a particular action to be a condition of moral responsibility for that action. What's required instead is something much closer to Frankfurt's identificationist axiom:

> (4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).

The attempt to use mechanisms as an extra layer between the action and the agent, mediating her responsibility, fails in several respects. Even if we set aside the criticisms I made in section 6.6, mechanisms cannot – at least in the way they are used by Fischer and Ravizza – establish the relevant ownership relationship between the agent and her action.

---

[94] I'm not going to try to give such an account. Indeed in the end I'm sceptical that any principled account of taking responsibility could be given: I suspect that an accurate account of taking responsibility would involve an arbitrary list of those kinds of actions for which we do in fact attribute responsibility.

[95] Such cases can involve for example brainwashing, indoctrination, hypnosis or direct stimulation of the brain.

[96] As Fischer and Ravizza explain in the first quotation above, taking responsibility for a kind of mechanism is "shorthand" for taking responsibility for the actions which issue from it.

An agent's counterfactual rejection of responsibility for an action, if she knew all the action's causes, should lead us to *excuse her from DMR* for that action. It should certainly weigh more heavily than the fact that she has taken responsibility for – and ownership of – the mechanism-kind which issued in the action. In the counterfactual scenario she has more information, which is very relevant to the issue in question; and that is the only difference between the counterfactual scenario and the actual situation.

I conclude that the Fischer-Ravizza guidance control conditions are not sufficient for direct moral responsibility. Though the guidance control conditions are met in the five key cases, each agent would reject responsibility for her action, if she knew all of its causes. Therefore we should not hold those agents directly morally responsible for their actions. I'll expand on this argument by considering some objections to it.

An important objection would make the intuitively appealing claim that agents cannot be excused responsibility for some of their actions on a piecemeal basis. Having taken responsibility for a mechanism-kind, the objector argues, an agent should not be excused responsibility for a certain action simply because she would prefer to reject that responsibility. Taking responsibility for – and ownership of – a mechanism-kind is a very important matter, and we must accept the consequent responsibilities which follow.

There is an intuitive merit in this way of thinking, but I think the problem of subverting causes reveals its limitations. It might be reasonable to expect agents to accept moral responsibility for all the actions which issue from a mechanism-kind, if those agents were aware of all the kinds of actions which might issue from it, when they took responsibility for the mechanism-kind. But that's exactly what we are not aware of. Assuming for the sake of argument that we do take responsibility for mechanism-kinds, there is a very significant flaw in the everyday model of action explanation which we use when we do so.[97] If we realised that actions with subverting causes might issue from those mechanism-kinds, we would not take responsibility for them, and all the actions which issue from them. It is therefore unreasonable to hold us morally responsible for all the actions which issue from a mechanism-kind. When we took responsibility for that mechanism-kind, we did so on the basis of a very significant misunderstanding of its operation.[98]

An obvious objection to this line of argument is that many agents *would* take responsibility for their mechanism-kind "ordinary practical reasoning", and hence the actions that issue from it, *even if* they knew that some of the actions which would issue from it would have subverting causes.

In response I accept that many people would do so, if they were simply told about a few example cases of subverting causes and asked for a prompt reaction to their new knowledge. But I think that, if they spent some time reflecting carefully on the implications of the wide-ranging research I discussed in sections 3.6, 3.7 and 3.8, most people would take the opposite view. Our intuitions about moral responsibility are closely linked to our everyday model of action explanation. Only after careful consideration of the evidence is someone likely to countenance a change to her model, to include serious consideration that some actions have subverting causes. Only after further reflection on the implications of this change will she then change her views about responsibility attribution. The fact that many people would initially retain their current views about responsibility, faced with some evidence about actions with subverting causes, does not show that they would retain those views after further reflection.

[97] By "the everyday model of action explanation", I mean the methods and tendencies used by non-experts to explain actions. Although there is of course no single uniform model, I take it that almost everyone ignores subverting causes when explaining almost every action. In sections 3.6, 3.7 and 3.8, I discussed evidence that this is a flaw in our everyday model of action explanation.

[98] Of course, I deny that we ever do take responsibility for mechanism-kinds; I am setting that point aside here.

A separate objection would be that agents can be DMR for actions with subverting causes in virtue of being (directly) *blameworthy* for performing them. I discussed several forms of this objection in section 4.3. As well as responding directly to the objection, I argued that Frankfurt would not raise it, because he holds axiom (4) above: an agent is DMR for an action only if it is truly the agent's own. Since Fischer and Ravizza do not hold this axiom, might they raise this objection?

I think they would not, because they do believe there must be an ownership relation between the agent and the action for which she is DMR. This relation is an important part of their strategy for dealing with cases of manipulation, brainwashing, and so forth, as well as Frankfurt-style cases. Fischer and Ravizza would not, I take it, argue that an agent can be directly blameworthy for an action if the ownership relation does not obtain. In their case, the ownership relation involves taking responsibility for actions and mechanism-kinds. I've argued that their account of the relation is flawed, and specifically that it breaks down in cases of actions with subverting causes. If they accepted those criticisms, Fischer and Ravizza would not press the present objection.

I need not rely only on this reply, though, since I did answer the objection directly (in several forms) in section 4.3.

It's possible that the problem I've highlighted for the Fischer-Ravizza theory could be solved by a minor revision. Rather than using "ordinary practical reasoning" as their primary example of a kind of mechanism, the authors might claim that there is a distinction to be drawn among kinds of mechanism. The suggestion would be that actions with subverting causes issue from a certain kind (or kinds) of mechanism, while other actions (including those we think of as paradigmatic examples of morally responsible action) issue from another kind (or kinds) of mechanism. On this view, agents take responsibility for the latter kind(s) of mechanism, and would endorse that judgement even if they learned about potential subverting causes of actions.

However, it seems very unlikely to me that this suggestion would find empirical support in either neuroscience or psychology. Instead, the *processes* which precede actions with subverting causes seem likely to be very similar to those which precede paradigmatically rational actions. For example, all of our actions involve a great many automatic processes of which we are not consciously aware. There seems no reason to think that the processes differ when actions have subverting rather than non-subverting causes. There's no obvious basis, then, for claiming that the *mechanism-kinds* differ when there are, or are not, subverting causes of actions.

6.8. CONCLUSIONS FROM THIS CHAPTER

In this chapter I've argued that there are flaws in all three elements of Fischer and Ravizza's theory.

Fischer and Ravizza's notion of "*the mechanism issuing in action*" is flawed. They make appeals to intuitions to support their claims about mechanisms, and also to support some of their claims about responsibility for actions performed by addicts and compulsives. Several of the most important appeals fail.

The Fischer-Ravizza conditions for "*moderate reasons-responsiveness*" include a requirement of "weak reactivity". This is simply too weak. If there is a reactivity condition for

moral responsibility, it must require more than that a mechanism would respond differently to a very extreme reason such as the threat of murder of the agent's children.[99]

Finally, the Fischer-Ravizza account of mechanism *ownership* is flawed. This is bound to be the case since the notions of mechanisms and kinds of mechanism are themselves flawed. But even setting aside those issues, the ownership requirements fail to ensure that the agent is morally responsible for every action which issues from a kind of mechanism which is owned.

There are two kinds of counterexample. The first involves agents acting in the grip of compulsions or addictions. Here it seems plausible to say that the mechanism issuing in action is not owned by the agent, since he has not taken responsibility for it.[100] But Fischer and Ravizza maintain, implausibly, that the mechanism is owned and the agent is morally responsible for his action.

The second kind of counterexample involves my five key cases of actions with subverting causes. In these it is plausible to agree with Fischer and Ravizza that the operative mechanism is owned, on their account of mechanism ownership. But this does not entail that the agents are morally responsible for their actions, even if the other conditions of guidance control are met. For if those agents knew all the causes of those actions, they would *reject responsibility* for the actions. This should lead us to conclude that those agents *are not directly morally responsible* for their actions.

Therefore, ownership of a moderately reasons-responsive mechanism which issues in an action – guidance control – is not sufficient for direct moral responsibility for that action.

[99] I think this problem is independent of the problems I've identified with the notion of the mechanism. A requirement of *weak reactivity of the agent*, rather than weak reactivity of the mechanism, would still be too weak. A plausible reactivity condition would require more than that *an agent* would react differently to a very extreme reason. (I haven't developed this point, as it isn't central to my argument.)

[100] On this way of applying the Fischer-Ravizza theory, the agent would not meet the guidance control conditions, and so would not be held directly morally responsible for the action – which is the plausible verdict.

7. THE PREVALENCE OF MORAL RESPONSIBILITY

In this chapter I'll make the following "sceptical argument". There are many everyday actions which have subverting causes. Many of those are actions for which we currently hold agents morally responsible. But, in many of those same actions, the agents are not in fact morally responsible – they bear neither direct nor traced moral responsibility.[1] I conclude that there are many everyday actions for which we mistakenly hold agents morally responsible.

This argument might be resisted by any of the compatibilists whose accounts I have discussed in previous chapters. For simplicity, I'll express my sceptical argument using the central elements of Frankfurt's theory of moral responsibility. Alternative versions could also be expressed using elements of the theories of Doris, Nahmias or Fischer and Ravizza: these alternative versions would have the same structure, and their conclusions would be the same.

The following is a very brief recap of my arguments against Frankfurt. On Frankfurt's final account, an agent identifies with his effective desire if and only if he accepts it as his own, and is satisfied with that acceptance, where being satisfied is a matter of having no interest in making changes.[2] I used the label "F-acceptance" to stand for this kind of acceptance. Frankfurt regards F-acceptance of an effective desire as sufficient for the desire's being truly one's own, and hence for direct moral responsibility for the action.[3] I argued against both of these points, using example cases of actions with subverting causes.[4] In these cases the agents do F-accept their effective desires, and yet the desires are not truly their own. I also made a positive proposal, that my CFA condition is necessary for direct moral responsibility:

> *CFA* The agent *would F-accept his action, if he knew all the proximate and relevant causes* of the action.[5]

I'll make use of these definitions in this chapter.

I'll make the case for my sceptical conclusion in section 7.1. Then, in section 7.2, I'll discuss whether research into subverting causes of actions also gives some grounds for optimism about moral responsibility. In section 7.3 I'll discuss some of the implications of my conclusions. My focus in this chapter is on actions with subverting causes, where those subverting causes are *unknown* both to the agents and to onlookers who may attribute moral responsibility. Therefore most of my examples will be of this kind of action.

7.1. MORAL RESPONSIBILITY FOR EVERYDAY ACTIONS

There are three parts to my sceptical argument. In section 7.1.1 I'll review support for the claim that there are many everyday actions with subverting causes. Then in section 7.1.2 I'll argue that agents are not morally responsible in many everyday actions with subverting causes. Finally in section 7.1.3 I'll argue that we do in fact hold people morally responsible for many of those actions. It follows that we mistakenly hold people morally responsible for many everyday actions.

---

[1] Direct moral responsibility (DMR) is anchored in facts about the action and the agent at the time of acting. Traced moral responsibility (TMR) can be "traced back" to facts about an earlier action (or omission) and about the agent at that earlier time. See section 1.2.

[2] See section 2.1.3.

[3] These are statements (5) and (6) in the set defining Frankfurt's position in section 2.2.

[4] See section 3.4.

[5] This is a simplified version of the condition; see section 3.5.

7.1.1. There are many actions with subverting causes

In chapter 3 I made these claims:

(A) There are actions with subverting causes.
(B) There are many everyday actions with subverting causes.

Claim (A) supports a key premise in my "main arguments" against Frankfurt and Fischer.[6] For my "sceptical argument", in this current chapter, I'll focus on claim (B). To establish that (B) is plausible, I defended these subsidiary claims:

(B1) There are good grounds to believe, on the basis of existing experimental evidence, that many everyday actions have subverting causes.
(B2) There are good grounds to believe that there are many more subverting causes of everyday actions as yet undiscovered by experimenters.

The term "many" in these claims is deliberately loose, and indicates a range of possibilities. Someone who finds it plausible that there are *very* many actions with subverting causes will find that my argument leads to a very sceptical conclusion.

To support claim (B1), I reviewed experimental evidence from several fields of psychology. There have been many studies of real-life everyday actions, or actions that are just like everyday actions, which have surprising causes.[7] It's very likely that many of those causes were subverting: the agents would not F-accept their actions, if they knew all the causes.[8] I also rejected some objections to these points.[9] Nevertheless, it is difficult to believe that everyday actions can have subverting causes. I discussed evidence which suggests explanations for our failing to notice that everyday actions have subverting causes.[10]

To support claim (B2), I gave the following inductive argument. Most of the evidence which supports claim (B1) has only recently been discovered. Almost all of the evidence dates from the last fifty years, and much of the most interesting evidence has come in the last decade. A lot of the evidence has been extremely surprising: we did not expect to discover that our actions can be influenced by the factors which have been pinpointed. Some of the research fields from which the evidence has come are still in their infancy, and the experimental methods used are steadily becoming more and more sophisticated. Therefore it is very likely that many more subverting causes and effects will be discovered by experimenters in the coming years.

Claims (B1) and (B2) are interrelated. Both are matters of degree. The more plausible is claim (B1), the less support is required from claim (B2), in order to establish (B). The more plausible is claim (B2), the less support is required from claim (B1).

7.1.2. There are many actions for which agents are not morally responsible

When an action has a subverting cause, the agent would not F-accept her action upon learning of its causes. I argued in chapter 3 that such an action is not truly the agent's own, and so she is not *directly* morally responsible for it. If there are many everyday actions which have subverting causes, then it follows that there are many everyday actions for which the agent is not directly morally responsible.

---

[6] See sections 3.4 and 6.7.

[7] One could argue that some studies of extremely unusual actions also provide evidence that similar effects operate in everyday actions. For example, some of the influencing effects of authority figures revealed in Milgram-style experiments may also occur in workplace meetings (Doris 2002:148). However, I don't need to rely on this point.

[8] See section 3.6.

[9] See section 3.7.

[10] I reviewed this evidence in section 3.8. I'll summarise it again in section 7.1.3 below.

I'll now argue, further, that there are many everyday actions with subverting causes for which the agent does not bear *traced* moral responsibility either. In order to bear TMR for an action with a subverting cause (which I'll call action A1), it must be the case that there was some earlier action or omission (A0) from which A1 results, and that the resultant occurrence of A1 was reasonably foreseeable by the agent at the time of A0.[11]

Now it seems clear that the effects of many subverting causes are not reasonably foreseeable by an agent who is not familiar with the relevant psychology research. Many of the effects discussed in chapter 3 were surprising even to psychologists when they were first discovered. They are surprising – even shocking – to all of us when we first learn about them. This is evidence enough that the effects are not foreseeable by a person who is unfamiliar with the psychology research.

To say that the effect of a subverting cause is not foreseeable by an agent is not quite the same as saying that the *resultant occurrence* of an action with a subverting cause was not reasonably foreseeable. But in fact we can see that the latter point does follow by considering a couple of examples.

Suppose that Michael is a subject in a Milgram-style study.[12] He agrees to take part in what he is told is a study of learning, and begins to follow the first few instructions given by the "experimenter" in the white coat. We might hope to trace moral responsibility for his later action to one of these earlier actions. Since Michael knows nothing about situationism, he cannot reasonably foresee that one or more situational factors will later *be causes of* his agreeing to administer a lethal dose of electricity to the "learner". It is not plausible to say that the *resultant occurrence of that action* is reasonably foreseeable by him at the time of the earlier actions. But perhaps that example is too far removed from everyday life to establish the point. The action of administering a lethal electric shock is extremely unusual, and it's difficult for anyone to foresee himself performing it. A more commonplace example may help.

Ingrid interrupts a conversation which was already in progress between two other people.[13] A cause of her action was that, a few minutes earlier, she completed a puzzle in which some of the words were connected with a theme of rudeness. It might be true that, at the time of completing the puzzle, Ingrid could have foreseen her later action of interrupting a conversation. But we need more than that to establish traced moral responsibility. Is it plausible to say that the *resultant occurrence* of her interrupting a conversation was reasonably foreseeable by her, at the time of completing the puzzle? It is not, because she does not know about the relevant psychology research which reveals this kind of effect. Therefore we cannot trace moral responsibility for the interrupting to the earlier action.

A separate point is that, in some cases, tracing of responsibility from an earlier action will be impossible because that earlier action itself had a subverting cause – and so the agent was not directly morally responsible for that earlier action.

It's also worth noting that if the tracing principle is false, then (if my proposed CFA condition is necessary for DMR) moral responsibility is impossible for actions with subverting causes.

I conclude that there are many everyday actions with subverting causes for which the agent bears neither direct nor traced moral responsibility.

[11] I discussed this definition of TMR in section 1.2. It seems plausible that the same argument would succeed on other definitions of TMR, too.
[12] See section 3.6.1.
[13] See section 3.2.4.

7.1.3. We mistakenly hold agents morally responsible for many actions

Most of the examples of subverting causes I've described are readily applicable to everyday actions. They involve people running late to give presentations, filling out paperwork, assessing job applicants, completing word puzzles, interrupting conversations, resisting temptations, and so on. Without knowledge of any subverting causes in these situations, we would readily hold the agents morally responsible for their actions. We would not even pause to doubt their moral responsibility. We simply do not expect that everyday actions like these might have subverting causes, and we do not allow for that possibility in our judgements about moral responsibility.

I have reviewed evidence which helps to explain why these examples of subverting causes are so surprising and conflict with common sense.[14] Situationist research reveals that our commonsense explanations of actions are much less reliable than we think. The phenomena of dissonance, confabulation and rationalisation all contribute to our failure to recognise mistakes in our explanations of our own actions. We lack awareness of the automatic processes involved in our actions: we are not consciously aware that they are occurring; nor do we realise how important their roles are. We have an experience of agency; but do not realise that it is interpreted automatically from factors such as the exclusivity and priority of our conscious thoughts about the action, and does not arise from introspective access to the processes which trigger and control actions. Furthermore, subverting causes can operate by distorting our *assessments* of situations in which we act: for example, we sometimes do not notice that we have failed to respond to a moral obligation.

The possibility of unknown subverting causes is not currently accounted for in everyday action explanation, except perhaps for such possible causes as subliminal perceptions or unconscious memories, which are present at the fringes of popular understanding.[15]

A problem we face as agents is that we don't know when our everyday actions have subverting causes. Suppose Delia is deliberating before choosing between action A and action B. In favour of A, she recognises reasons R1 and R2. In favour of B are reasons R3 and R4. She decides to perform action A. But before she does so, she changes her mind and instead performs action B. Now asked why she changed her mind, Delia says that she re-evaluated reason R3 as more important than she had previously thought. Unfortunately, Delia cannot know whether her action had a subverting cause. It may be that a subverting cause led to her change of mind, and she confabulated the re-evaluation of reason R3 (which did not happen). Alternatively, it may be that a subverting cause led her to re-evaluate R3: in this case, the explanation is correct but the action is not truly her own.[16] Delia knows that there were reasons in favour of action B, but does not know that there were no subverting causes. This epistemic limitation is problematic in itself, independently of the implications for moral responsibility for our own actions. It means that we cannot be confident in our judgements about other agents' actions, even when they give sincere explanations.[17]

It's impossible to calculate the proportion of everyday actions in which there are subverting causes, for several reasons. The fields of research into situational factors, and other factors processed automatically in unexpected ways, are still very young. We simply don't know how many surprising effects of these kinds still lie undiscovered. Furthermore, a potentially subverting factor will sometimes not subvert an action, either because it is not a cause of the action, or because the agent would F-accept the action if he knew of its effect.

[14] This paragraph summarises evidence discussed in section 3.8.

[15] I did not include examples of subliminal perception or unconscious memory among my five key cases, because I'm not aware of specific evidence which demonstrates that they are causes in everyday actions. It may well be, however, that such evidence exists or will soon be discovered.

[16] That is, if she learned of the (subverting) cause of her action, she would not F-accept it as her own.

[17] There are also implications for our view of ourselves as autonomous agents, which I won't pursue here.

Before moving on, I'll briefly summarise my sceptical argument. I think it's plausible that there are many actions with subverting causes for which the agent does not bear *traced* moral responsibility. This can be the case when the agent cannot reasonably foresee the operation of the subverting cause. It can also be the case when the prior action, from which responsibility might be traced, itself had a subverting cause. Meanwhile an agent cannot be *directly* morally responsible for an action with a subverting cause, because it is not truly his own. There are many everyday actions with subverting causes. Therefore there are many everyday actions for which agents are not morally responsible. Many of these are actions for which we currently do hold agents morally responsible, in practice, because we are not aware that they may have subverting causes. It follows that there are many everyday actions for which we mistakenly hold agents morally responsible.

7.2. GROUNDS FOR OPTIMISM ABOUT MORAL RESPONSIBILITY

An optimistic response to the challenges posed by research on subverting causes is to hope that increasing knowledge of the effects will help us in future to perform more actions which are truly our own. Both Nahmias and Doris emphasise the possible increases in freedom and responsibility to be gained by greater knowledge of situationist effects. There are three ways in which increased knowledge of subverting causes might help someone. First, she may be able to recognise potential subverting causes before they operate, so that they do not become causes of her actions. Second, the person may reduce her reliance on character traits – both as explanations of other people's actions, and more importantly as predictors of her own future actions. She may stop relying complacently on predictions of what she will do in some future situation which are based predominantly on character traits. Third, she may learn how to manipulate her environment so as to reduce either the prevalence or the influence of potential subverting causes. I'll briefly assess each of these possibilities in this section.

Unfortunately there has been little relevant research to establish whether the first source of optimism presents good grounds for hope.[18] We must be careful to avoid moving complacently from an unrealistic view of how our actions are caused to an unrealistic vision of what we can achieve in the future. It seems intuitively plausible, for example, that teaching someone about the 'Good Samaritan' experiment, or the effect of bystanders' behaviour on helping people in need, would have a significant on her future actions in any similar circumstances. But the (limited) experimental evidence suggests that any such effect is weak.[19] Doris speculates that better results could be gained by "integrating the lessons learned from situationism into our culture of moral education".[20] That seems to me much more likely to be effective than individuals learning fragments of knowledge about situationism and attempting to modify their actions. It would be worthwhile to include non-situationist subverting causes in this education too – such as factors processed automatically in unexpected ways.

The second optimistic hope is that agents may learn to reduce their reliance on character traits, both when predicting their own future actions and when explaining others' actions. I'm not aware of any experimental evidence to support this hope. It may be that attributing character traits and using them in these ways is instinctive. There may be some evolutionary

[18] Doris 2002:148.
[19] In one study, students who had attended a lecture or film about group effects and helping behaviour were more likely than control subjects to intervene two weeks later to help someone, when part of a group of bystanders. However, "the effect, although significant, is not extremely marked" (Doris 2002:148).
[20] Doris 2002:148.

advantage to our doing so.[21] To some extent this second optimistic hope is linked to the first: the difficulty of success in one is probably reflected in the other.

In my view the third optimistic hope is the one most likely to bring results in the short term. We can change our behaviour to *avoid* situations in which subverting causes may be prevalent. For example, with more understanding of situational pressures, Alf may choose to decline the invitation to a candlelit dinner which might lead him to a later akratic act of infidelity.[22] Manipulating our environment can enable us to reduce the prevalence or the influence of potential subverting causes. We may also be able to introduce factors which subtly incline us *toward* behaviour that we want to encourage, making use of research on situationism and on factors processed automatically in unexpected ways.[23] A problem for this line of optimism is that we may need to take such steps *very* often in future: trying to avoid the influence of *many* potentially subverting factors may require a lot more effort than trying to recognise individual factors.

There are practical problems with all three of these optimistic responses. One problem is that it will be difficult for psychologists to uncover every type of subverting cause. Until they do – or if they do not succeed – we should avoid overconfidence in our efforts to overcome such causes. We will continue to face the prospect that there may be more unknown factors that will subvert our actions.

Perhaps even more problematic is the possibility that there are simply *too many* potentially subverting factors. Even if they were all studied and documented by scientists, it might not be practical to use all the relevant available knowledge in an everyday action. Just trying to take into account the few kinds of effect I have discussed here would be taxing, and perhaps impractically difficult. For example, if Charlotte decides to offer a job to a candidate, she might need to ponder whether this action could have been subverted by: running late; other people's behaviour; recently having processed any words which might affect her mood; recently resisting some temptation; recently engaging in a logical thinking activity; whether the candidate spilled coffee at her interview; and so on. This list already seems too long to be manageable, even if Charlotte made a conscious effort to consider all these possibilities. If she did not make the effort, she might well fail to notice factors which have the potential to subvert her action. And since there are many more known factors which I haven't discussed or listed here, the practical problem seems unmanageable. And that problem will only be made worse by a growing number of discoveries from future research.

This practical problem also has implications for moral responsibility. The sheer number of possible subverting effects is so large that it seems implausible to hold someone morally responsible if she failed to avoid all of them – even if she did know about all of them. If a certain factor was a subverting cause of a person's action, then the action was not truly her own, and so she is not directly morally responsible for it. And even if she had knowledge about that particular type of subverting effect and had an opportunity to avert it by some earlier action or omission, it seems implausible to say that her subverted action's resulting from that earlier action was reasonably foreseeable, if there were so many potentially subverting factors that she simply could not consider them all. Therefore it may be impossible to attribute TMR to the agent even if she did know about the subverting effect which was a cause of her action.

[21] It may be that action explanation in terms of character traits facilitates clear-cut and widely accepted judgements of responsibility. This in turn may strengthen the effect of those judgements in deterring antisocial actions.

[22] See section 5.1.

[23] See Thaler and Sunstein 2008 for many interesting examples of how we may be "nudged" towards preferable behaviour by manipulated environmental factors – whether manipulated by ourselves or by others.

Setting aside all of those problems, the main limitation of all three optimistic responses to subverting causes is that they do not constitute objections to my argument. By the three means suggested we may be able, *in future*, to perform more everyday actions which are truly our own. But that does not challenge my conclusion that there are currently many everyday actions which are not truly our own, and for which we are not directly morally responsible. Nor do they establish a means by which we can happily accept traced moral responsibility for actions with subverting causes. In fact, we don't *want* to bear TMR for actions which have subverting causes, since those actions are not truly our own.[24]

My conclusion remains, notwithstanding the possible grounds for future optimism, that today there are many everyday actions for which agents are mistakenly held morally responsible.

7.3. IMPLICATIONS OF MY CONCLUSIONS

I have argued against two of the most influential contemporary compatibilist theories of moral responsibility. They are also the leading representatives of two of the most important *kinds* of compatibilist theory: Frankfurt's is the most influential identificationist theory of responsibility, while the Fischer-Ravizza account is the most widely-discussed and detailed attempt to analyse responsibility in terms of responsiveness to reasons. I have also considered and rejected two attempts (by Doris and Nahmias) to build identificationist theories which take account of the threats to moral responsibility posed by research in situationist social psychology.[25] I've argued that none of these theories provides conditions which are sufficient for moral responsibility. My arguments employ the central elements from Frankfurt's and Fischer's own positions, so I believe that they should be accepted by adherents of those positions.

In this chapter I've also argued for a very sceptical conclusion: there are many everyday actions for which agents are not morally responsible.

Determinism plays no role in any of my arguments. The truth or falsity of determinism is irrelevant to the truth or falsity of my conclusions. Both Frankfurt and Fischer reject the Principle of Alternate Possibilities (PAP), according to which "a person is morally responsible for what he has done only if he could have done otherwise".[26] By this means they defend their theories from the most common arguments for the incompatibility of determinism and moral responsibility. Very broadly, these arguments run as follows: if PAP is true, and if determinism is incompatible with alternate possibilities, then determinism is incompatible with moral responsibility.[27]

Questions about determinism and its implications are so important and far-reaching that they continue to dominate discussion of free will and moral responsibility. For example, Fischer declares:

> I am motivated in much of my work by the idea that our basic status as distinctively free and morally responsible agents should not depend on the arcane ruminations – and deliverances – of the theoretical physicists and

[24] It may sometimes seem otherwise. But if, for example, Alf accepts the invitation to a candlelit dinner despite *knowing* that it may lead him to an akratic act of infidelity, then that later act probably *is* truly his own (see section 5.1).

[25] To my knowledge these are the only theories of moral responsibility which address the threats posed by subverting causes.

[26] Frankfurt 1969:1.

[27] That determinism is incompatible with alternate possibilities is the conclusion of the family of arguments known as "the Consequence Argument" (of which the most influential versions are found in van Inwagen 1983). Some compatibilists accept PAP but deny that determinism is incompatible with alternate possibilities.

> cosmologists. That is, I do not think our status as morally responsible persons should depend on whether or not causal determinism is true.[28]

The arguments I've presented here pose a threat to the theories of Frankfurt and Fischer from a very different direction. I've argued that their proposed conditions are not sufficient for direct moral responsibility, *even if* PAP is false, and *whether or not* determinism is true and incompatible with PAP. The source of the threat is not theoretical physics but empirical psychology. And yet the sceptical conclusion I've defended is extremely significant: we mistakenly hold agents morally responsible for many everyday actions.[29]

If all of my conclusions are correct, then followers of Frankfurt and Fischer face the following options. First, they could accept that there are many everyday actions for which agents are not morally responsible. This is an unpalatable consequence, which compatibilists will surely resist.

Second, they could adopt a different compatibilist account of the same kind. Followers of Frankfurt might adopt an alternative form of identificationism; followers of Fischer might support or develop another theory in which moral responsibility is analysed in terms of responsiveness to reasons.

I've discussed two identificationist theories which do take account of situationist research. I argued that neither Doris's nor Nahmias's conditions of identification are sufficient for direct moral responsibility. I think it's likely that other identificationist accounts will face similar difficulties – though I have not argued for this conclusion. I find it hard to imagine a plausible condition of identification on which the agent's effective desire, in the five key cases of actions with subverting causes, would be truly her own (in the sense which is necessary for DMR).[30]

Nor can I imagine a plausible set of conditions involving reasons-responsiveness which would be sufficient for DMR, and also satisfied in the five key cases. The problem, when an action has a subverting cause, is not that the agent (or mechanism) is not responding to reasons. Nor is the problem that the agent (or mechanism) is not responsive to reasons. The problem is something closer to this: *the agent is not responding to reasons in the way that she would wish to respond*, if she knew all the causes of her action.

I have not shown, of course, that no suitable account of moral responsibility can be given which would be appealing to followers of Frankfurt or Fischer. My objective here is simply to indicate that the task may be difficult.

A third option facing followers of Frankfurt and Fischer, if my conclusions are correct, is to adopt a different *kind* of compatibilist theory of moral responsibility – to move away from theories centred on identification and reasons-responsiveness. I haven't examined any such theories, but again I think it's likely that many compatibilist accounts of responsibility will come under attack as the phenomena of actions with subverting causes become more widely discussed. Few philosophers have addressed the threats posed to moral responsibility by *situationist* research.[31] Recent research in other fields of psychology reveals that there are many more types of subverting cause, of which word puzzles and resisted cookies are examples.[32] But in any case, followers of Frankfurt will not want to abandon identificationism

[28] Fischer 2006:5.

[29] My arguments give no support to incompatibilist libertarians, either – except in so much as they highlight problems for these influential compatibilist theories. The problems posed by actions with subverting causes may threaten libertarian attempts to define the sufficient conditions for direct moral responsibility in similar ways.

[30] I argued that my CFA condition, which is based on Frankfurt's identification condition, is necessary for DMR. But the CFA condition is not met in actions with subverting causes. (See section 3.5.)

[31] To my knowledge the only philosophers who have done so are Doris, Nahmias and Nelkin (2005).

[32] For more examples see section 3.6.2.

altogether; nor will followers of Fischer want to endorse theories which do not analyse moral responsibility in terms of reasons-responsiveness.

I can see no readily available option which would be palatable to followers of Frankfurt or Fischer, if they accept my criticisms. It's likely that they would want to refute my arguments, rather than pursue one of these options. This shows that my conclusions are significant, if correct.

7.4. CONCLUSIONS

For my "sceptical argument" in this chapter, I drew on my earlier arguments (given in chapter 3) that there are many everyday actions which have subverting causes, and that agents are not directly morally responsible for actions which have subverting causes. I showed that many of those are actions for which we currently hold agents morally responsible. I also argued that the agents of many of those actions do not bear traced moral responsibility for them. Therefore I conclude that we mistakenly hold agents morally responsible for many everyday actions.

Although this conclusion is troubling, and may seem unduly sceptical, it's worth noting that I rejected some compelling objections which would have *even more* sceptical implications. My line on non-subverting situationist causes – such as finding a dime in a phone booth before acting to help a stranger – faced several objections which led toward more sceptical conclusions.[33] I also rejected Nahmias's conclusion that two conditions are necessary for direct moral responsibility: his Knowledge condition and Counterfactual Awareness condition.[34] In my view, either of these – if necessary for DMR – would entail a smaller number of actions for which agents do bear direct moral responsibility. None of these alternative and more sceptical views is completely implausible; I was hesitant to reject each of them. Although the conclusion I reach is seemingly extreme, there are more extreme alternatives.

My sceptical conclusion, if correct, is obviously significant. The prospect that there are many everyday actions for which agents are not morally responsible is unpalatable to compatibilists. Followers of Frankfurt and Fischer will surely object to my arguments, rather than accept this conclusion.

[33] See section 4.1.
[34] See section 5.2.1.

8. SUMMARY OF MY ARGUMENTS AND CONCLUSIONS

In this final chapter I will summarise my arguments and conclusions. I will focus on the arguments which contribute most to my main objectives; many points and arguments from earlier chapters won't be included in this summary.

I began in chapter 2 by describing the development of Frankfurt's identificationist theory of moral responsibility. On Frankfurt's final account, an agent identifies with an effective desire to act if he *accepts* the desire as his own, and if he is *satisfied* with that attitude of acceptance, where being satisfied is a matter of having no interest in making changes. I introduced the term "F-acceptance" to stand for this kind of acceptance.

I summarised Frankfurt's final position in a set of statements defining the relationships between four states of affairs:

(A) The agent *F-accepts* his effective desire.
(I) The agent *is identified with* his effective desire.
(O) The agent's effective desire *is truly his own*.
(R) The agent is *directly morally responsible* for the action to which he is moved by his effective desire.[1]

The six statements are:

(1) F-acceptance is necessary and sufficient for identification (A↔I).
(2) Identification with an effective desire is sufficient for its being truly the agent's own (I→O).
(3) Identification is sufficient for direct moral responsibility (I→R)
(4) For direct moral responsibility for an action, the effective desire must be truly the agent's own (R→O).
(5) F-acceptance of an effective desire is sufficient for its being truly the agent's own (A→O).
(6) F-acceptance of the effective desire is sufficient for direct moral responsibility for the action (A→R).

In my "main argument" against Frankfurt, in chapter 3, I argued that statement (5) is false: F-acceptance of an effective desire is *not* sufficient for its being truly the agent's own. I drew on five "key cases" of actions with subverting causes.[2] An action has subverting causes when it is true that: *if the agent knew all the proximate and relevant causes of her action, she would not F-accept the effective desire*.

I argued that, in each of the five key cases, the agent does F-accept her effective desire, and yet that effective desire is not truly her own. In each case, the agent would not F-accept her action, if she knew all of its causes. In this counterfactual scenario, the agent has more knowledge than she has in the actual situation, and that knowledge is relevant to her F-acceptance. Her counterfactual lack of F-acceptance indicates that the effective desire is not truly her own. Thus statement (5) is false: F-acceptance of an effective desire is not sufficient for its being truly the agent's own.

[1] By *direct* moral responsibility I mean moral responsibility which is not *traced* back to facts about an earlier action or omission. See section 1.2.

[2] As a brief reminder, the five key cases of actions with subverting causes on which I built my main arguments were the following. Samantha walks past a man slumped in doorway after being told that she is running late for her appointment to give a presentation. Martha, like the person sitting next to her, continues to fill out a market research questionnaire despite loud cries from the researcher in the next room. Judy judges that a candidate possesses a high degree of flexibility in solving problems, after being told that she will meet the candidate later. Ingrid interrupts a conversation shortly after completing a puzzle in which some of the words were connected with a theme of rudeness. Gina gives up trying to solve a difficult problem, after earlier resisting the temptation to eat some chocolate cookies. (See section 3.2 for full details.)

If statement (5) is false, it follows that (6) is false if the identificationist axiom (4) is true. Frankfurt's condition of identification, the agent's F-acceptance of his effective desire, is not sufficient for direct moral responsibility for the action.

I also made a positive proposal, that the following condition is necessary for direct moral responsibility:

> *CFA* The agent would F-accept the effective desire if she knew all the proximate and relevant causes of the action.[3]

I think followers of Frankfurt should accept my argument, since it employs the central elements from his own theory in a way which is consistent with the theory's objectives.

In the second half of chapter 3 I reviewed experimental evidence which shows that the five key cases of actions with subverting causes are not isolated incidents. Rather, the cases are representative of a large body of credible research in mainstream fields of psychology. I also reviewed evidence which explains why it is difficult for us to believe that there are actions with subverting causes. I addressed objections to my arguments against Frankfurt in chapter 4.[4]

In chapter 5 I discussed two philosophers – John Doris and Eddy Nahmias – who have developed identificationist theories of moral responsibility specifically in response to the threats posed by *situationist* subverting causes.[5] I argued that Doris' "Narrative Integration" condition would entail implausible revisions to our practices of responsibility attribution. Furthermore, it's not clear that judgements about narrative integration are genuinely independent of judgements about moral responsibility. Meanwhile Doris's treatment of actions with subverting situationist causes is problematic. He claims that knowledge of situational factors facilitates increased responsibility, but a lack of such knowledge is not exculpating. These claims seem to be in tension with one another.

I also argued, using counterexamples, that Doris' narrative integration condition is neither necessary nor sufficient for direct moral responsibility.

Nahmias proposes a "Knowledge condition" and what I called a "Counterfactual Awareness condition", both of which he takes to be necessary for direct moral responsibility. I argued that the Knowledge condition is too stringent: it's not plausible to believe that it is necessary for DMR. Meanwhile a modified version of my CFA condition, which incorporates a generic definition of identification, is preferable to Nahmias's Counterfactual Awareness condition as a necessary condition of direct moral responsibility:

> *CFG* The agent would embrace her determinative motive or regard it as fully her own, if she knew all the proximate and relevant *causes* of the action.

Using an example action with a situationist subverting cause, I also argued that Nahmias's Knowledge and Counterfactual Awareness conditions are not jointly sufficient for direct moral responsibility.

John Martin Fischer, working with Mark Ravizza, has developed the most influential example of a very different kind of theory of moral responsibility. On their account, moral responsibility is to be analysed in terms of reasons-responsiveness. I argued that there are problems with all three of the central elements of the Fischer-Ravizza theory: the notion of mechanisms, ownership of mechanisms, and the requirement of "moderate reasons-

[3] This is a simplified version of the condition; for the full version see section 3.5.

[4] Very similar objections could be made to my arguments in later chapters, and my replies, too, would be very similar.

[5] The first three of my five key cases are examples of situationist subverting causes of actions.

responsiveness".[6] These problems are exposed in the theory's treatment of agents suffering addictions and compulsions; but I set those cases aside for the sake of my main argument, which concerns actions with subverting causes.

The Fischer-Ravizza guidance control conditions are met in the five key cases of actions with subverting causes. In each case, the mechanisms are moderately reasons-responsive. The mechanisms are of the kind "ordinary practical reasoning", for which the agents have taken responsibility – and ownership – in virtue of having taken responsibility for past actions issuing from "ordinary practical reasoning".

I argued that, if each agent knew all the causes of her action, *she would reject responsibility for her action*. She would not accept that she is a fair target of reactive attitudes in relation to the action, if she learned of its subverting causes. This should lead us to excuse the agent from direct moral responsibility for the action. Thus Fischer and Ravizza's guidance control conditions are not sufficient for direct moral responsibility.

Ownership of a kind of mechanism is too inflexible to be a condition of moral responsibility for actions: it is too far removed from the causes of individual actions. This problem is already evident in cases involving addictions and compulsions, but is revealed even more starkly by the phenomena of subverting causes of actions.

Finally in chapter 7 I presented a "sceptical argument". There are good grounds to believe, on the basis of existing experimental evidence, that many everyday actions have subverting causes. There are also good grounds to believe that there are many more subverting causes of everyday actions as yet undiscovered by experimenters. So there are good grounds to believe that many everyday actions have subverting causes. Many of those are actions for which we currently hold agents morally responsible. But, as I argued in earlier chapters, agents are not directly morally responsible for actions which have subverting causes. Furthermore, there are many everyday actions with subverting causes for which the agent does not bear traced moral responsibility either, because the resultant occurrence of the subverted action is not reasonably foreseeable to her at the time of any earlier action or omission. Therefore, there are many everyday actions for which we mistakenly hold agents morally responsible.

Meanwhile, there are some grounds for optimism that increasing knowledge of the effects will help us to perform more actions which are truly our own. We may in future be able to recognise potential subverting causes before they operate, improve our predictions and explanations of actions, and learn how to manipulate our environments so as to reduce either the prevalence or the influence of potential subverting causes. Unfortunately, there is little experimental evidence so far to establish whether these are *good* grounds for optimism. And, in any case, the sheer quantity of potentially subverting factors in our environments imposes severe practical difficulties: it's not clear that an individual can learn about a large number of such effects, nor that she can take the steps required to mitigate their influence.

At the same time, these possible grounds for future optimism do not challenge my sceptical conclusion: there are currently many everyday actions for which agents are not morally responsible.

[6] The problem lies in the "weak reactivity" condition, which forms part of the requirement of "moderate reasons-responsiveness".

BIBLIOGRAPHY

Anscombe, Elizabeth (1957/1963). *Intention*, second edition. Cambridge, MA: Harvard University Press.

Ayer, Alfred (1954). 'Freedom and Necessity' in *Free Will*, first edition, edited by Gary Watson. Oxford: Oxford University Press.

Bargh, John (1994). 'The Four Horsemen of Automaticity: Awareness, Intention, Efficiency and Control in Social Cognition' in *Handbook of Social Cognition Volume 1: Basic Processes*, second edition, edited by Robert Wyer and Thomas Srull. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bargh, John (2005). 'Bypassing the Will: Toward Demystifying the Nonconscious Control of Social Behavior' in *The New Unconscious (Oxford Series in Social Cognition and Social Neuroscience)*, edited by Ran Hassin (et al.). Oxford: Oxford University Press.

Bargh, John (2008). 'Free Will is Un-natural' in *Are We Free?: Psychology and Free Will*, edited by John Baer (et al.). Oxford: Oxford University Press.

Bargh, John; Chen, Mark and Burrows, Lara (1996). 'Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action' in *The Journal of Personality and Social Psychology* 71/2:230-244.

Baumeister, Roy (2005). *The Cultural Animal: Human Nature, Meaning, and Social Life*. Oxford: Oxford University Press.

Baumeister, Roy (2008). 'Free Will, Consciousness and Cultural Animals' in *Are We Free?: Psychology and Free Will*, edited by John Baer (et al.). Oxford: Oxford University Press.

Baumeister, Roy; Blatslavsky, Ellen; Muraven, Mark and Tice, Dianne (1998). 'Ego Depletion: Is the Active Self a Limited Resource?' in *The Journal of Personality and Social Psychology* 74/5:1252-1265.

Berofsky, Bernard (2002). 'Ifs, Cans and Free Will: The Issues' in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.

Bratman, Michael (1996). 'Identification, Decision, and Treating as a Reason' in *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Bratman, Michael (2000). 'Fischer and Ravizza on Moral Responsibility and History' in *Philosophy and Phenomenological Research* 61/2:453-458.

Bratman, Michael (2002a). 'Hierarchy, Circularity, and Double Reduction' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.

Bratman, Michael (2002b). 'Nozick on Free Will' in *Structures of Agency: Essays*. New York: Oxford University Press.

Bratman, Michael (2006). 'A Thoughtful and Reasonable Stability' in *Taking Ourselves Seriously and Getting It Right*: *The Tanner Lectures in Moral Philosophy* by Harry Frankfurt, edited by Debra Satz. Stanford: Stanford University Press.

Buss, Sarah and Overton, Lee (2002). 'Introduction' in *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, MA: MIT Press.

Chandler, Teresa (2004). *Identification and Autonomy: A Meditation on the Philosophy of Harry Frankfurt*. Doctoral dissertation, University of Maryland.

Chisholm, Roderick (1964). 'Human Freedom and the Self' in *Free Will*, second edition, edited by Gary Watson. Oxford: Oxford University Press.

Clark, Andy (2007). 'Soft Selves and Ecological Control' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.

Dan-Cohen, Meir (2006). 'Socializing Harry' in *Taking Ourselves Seriously and Getting It Right*: *The Tanner Lectures in Moral Philosophy* by Harry Frankfurt, edited by Debra Satz. Stanford: Stanford University Press.

Darley, John and Batson, Daniel (1973). 'From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior' in *The Journal of Personality and Social Psychology* 27/1:100-108.

Davidson, Donald (1963). 'Actions, Reasons, and Causes' in *Essays on Actions and Events*, second edition. Oxford: Oxford University Press.

Davidson, Donald (1969). 'How is Weakness of the Will Possible?' in *Essays on Actions and Events*, second edition. Oxford: Oxford University Press.

Davidson, Donald (1971). 'Agency' in *Essays on Actions and Events*, second edition. Oxford: Oxford University Press.

Davidson, Donald (1973). 'Freedom to Act' in *Essays on Actions and Events*, second edition. Oxford: Oxford University Press.

Dennett, Daniel (1973). 'Mechanism and Responsibility' in *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton: Harvester Press.

Dennett, Daniel (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* Oxford: Oxford University Press.

Dennett, Daniel (1991). *Consciousness Explained*. London: Penguin Books.

Dennett, Daniel (1992). 'The Self as a Center of Narrative Gravity'. URL= <http://ase.tufts.edu/cogstud/papers/selfctr.htm>. Accessed 14th October 2005.

Dennett, Daniel (2003a). *Freedom Evolves.* London: Penguin Books.

Dennett, Daniel (2003b). 'Making Ourselves At Home in Our Machines (Review of Daniel Wegner's *The Illusion of Conscious Will*)' URL= <http://ase.tufts.edu/cogstud/papers/wagnerreview.htm>. Accessed 14th October 2005.

Dennett, Daniel (2004). 'Calling In the Cartesian Loans' in *Behavioral and Brain Sciences* 27/5:661-661.

Dennett, Daniel (2005). 'Natural Freedom' in *Metaphilosophy* 36/4:449-459.

Dennett, Daniel (2007). 'My Body Has a Mind of Its Own' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.

Dennett, Daniel (2008). 'Some Observations on the Psychology of Thinking About Free Will' in *Are We Free?: Psychology and Free Will*, edited by John Baer (et al.). Oxford: Oxford University Press.

Dijksterhuis, Ap; Aarts, Henk and Smith, Pamela (2005). 'The Power of the Subliminal: On Subliminal Persuasion and Other Potential Applications' in *The New Unconscious (Oxford Series in Social Cognition and Social Neuroscience)*, edited by Ran Hassin (et al.). Oxford: Oxford University Press.

Dijksterhuis, Ap; Smith, Pamela; van Baaren, Rick and Wigboldus, Daniël (2005). 'The Unconscious Consumer: Effects of Environment on Consumer Behavior' in *The Journal of Consumer Psychology* 15/3:193-202.

Dixon, Norman (1981). *Preconscious Processing*. Chichester: John Wiley & Sons.

Doris, John (2002). *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Doris, John (2005). 'Evidence and Sensibility (Symposium Replies)' in *Philosophy and Phenomenological Research* 71/3:656-677.

Dutton, Donald and Aron, Arthur (1974). 'Some Evidence for Heightened Sexual Attraction Under Conditions of Anxiety' in *The Journal of Personality and Social Psychology* 30/4:510-517.

Ekstrom, Laura (2005). 'Autonomy and Personal Integration' in *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*, edited by James Taylor. New York: Cambridge University Press.

Eshleman, Andrew (2001). 'Being is not Believing: Fischer and Ravizza on Taking Responsibility' in *The Australasian Journal of Philosophy* 79:479-90.

Eshleman, Andrew (2009). 'Moral Responsibility' in *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. URL = <http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/>. Accessed 28th May 2010.

Feinberg, Todd (2001). *Altered Egos: How the Brain Creates the Self*. New York: Oxford University Press.

Fischer, John (1987). 'Responsiveness and Moral Responsibility' in *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, John (1994). *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell.

Fischer, John (2002a). 'Frankfurt-Style Compatibilism' in *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, John (2002b). 'Frankfurt-type Examples and Semi-Compatibilism' in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.

Fischer, John (2003). 'Responsibility and Alternative Possibilities' in *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, John (2004). 'Responsibility and Manipulation' in *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, John (2006). 'A Framework for Moral Responsibility' in *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, John (2007). 'Compatibilism' in *Four Views on Free Will*, John Fischer (et al.). Oxford: Blackwell.

Fischer, John and Ravizza, Mark (1993). 'Introduction' in *Perspectives on Moral Responsibility*. New York: Cornell University Press.

Fischer, John and Ravizza, Mark (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, John and Ravizza, Mark (2000a). 'Précis of *Responsibility and Control: A Theory of Moral Responsibility*' in *Philosophy and Phenomenological Research* 61/2:441-445.

Fischer, John and Ravizza, Mark (2000b). 'Replies' in *Philosophy and Phenomenological Research* 61/2:467-480.

Fischer, John and Tognazzini, Neal (2009). 'The Truth About Tracing' in *Noûs* 43/3:531-556.

Flanagan, Owen (2002). *The Problem of the Soul: Two Visions of Mind and How to Reconcile Them*. New York: Perseus Books.

Frankfurt, Harry (1969). 'Alternate Possibilities and Moral Responsibility' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1971). 'Freedom of the Will and the Concept of a Person' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1975). 'Three Concepts of Free Action' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1976). 'Identification and Externality' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1978). 'The Problem of Action' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1987a). 'Identification and Wholeheartedness' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1987b). 'Rationality and the Unthinkable' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1988). 'Preface' in *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.

Frankfurt, Harry (1992). 'The Faintest Passion' in *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

Frankfurt, Harry (1993). 'On the Necessity of Ideals' in *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.
Frankfurt, Harry (1994). 'Autonomy, Necessity and Love' in *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.
Frankfurt, Harry (2002a). 'Reply to Eleonore Stump' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Frankfurt, Harry (2002b). 'Reply to Michael E. Bratman' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Frankfurt, Harry (2002c). 'Reply to Gary Watson' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Frankfurt, Harry (2002d). 'Reply to John Martin Fischer' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Frankfurt, Harry (2002e). 'Reply to J. David Velleman' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Frankfurt, Harry (2006a). 'Taking Ourselves Seriously' in *Taking Ourselves Seriously and Getting It Right*: *The Tanner Lectures in Moral Philosophy*, edited by Debra Satz. Stanford: Stanford University Press.
Frankfurt, Harry (2006b). 'Getting It Right' in *Taking Ourselves Seriously and Getting It Right*: *The Tanner Lectures in Moral Philosophy*, edited by Debra Satz. Stanford: Stanford University Press.
Ginet, Carl (1990). *On Action*. Cambridge: Cambridge University Press.
Haidt, Jonathan (2006). *The Happiness Hypothesis: Putting Ancient Wisdom to the Test of Modern Science*. London: William Heinemann.
Haji, Ishtiyaque (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.
Haji, Ishtiyaque (2002). 'Compatibilist Views of Freedom and Responsibility' in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.
Honderich, Ted (1988). *A Theory of Determinism, Volume 2: The Consequences of Determinism*. Oxford: Oxford University Press.
Hornsby, Jennifer (1980). 'Agency and Causal Explanation' in *The Philosophy of Action*, edited by Alfred Mele. Oxford: Oxford University Press.
Hume, David (1740/1978). *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, second edition. Oxford: Oxford University Press.
Hume, David (1777/1975). *Enquiries Concerning Human Understanding and the Principles of Morals*, edited by L. A. Selby-Bigge, third edition. Oxford: Oxford University Press.
Isen, Alice and Levin, Paula (1972). 'Effect of Feeling Good on Helping: Cookies and Kindness' in *The Journal of Personality and Social Psychology* 21/3:384-388.
Jack, Anthony and Robbins, Philip (2004). 'The Illusory Triumph of Machine Over Mind: Wegner's Eliminativism and the Real Promise of Psychology' in *Behavioral and Brain Sciences* 27/5:665-666.
Kane, Robert (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
Kane, Robert (2002a). 'The Contours of Contemporary Free Will Debates' in *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.
Kane, Robert (2002b). 'Some Neglected Pathways in the Free Will Labyrinth' in *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.

Kane, Robert (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.
Karremans, Johan; Stroebe, Wolfgang and Claus, Jasper (2006). 'Beyond Vicary's Fantasies: The Impact of Subliminal Priming and Brand Choice' in *The Journal of Experimental Social Psychology* 42:792-798.
Kihlstrom, John (2004). 'An Unwarrantable Impertinence' in *Behavioral and Brain Sciences* 27/5:666-667.
Knobe, Joshua (2006). 'The Concept of Intentional Action: A Case Study in the Use of Folk Psychology' in *Experimental Philosophy*, edited by Joshua Knobe and Shaun Nichols. Oxford: Oxford University Press.
Latané, Bibb and Darley, John (1970). *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts.
Latané, Bibb and Rodin, Judith (1969). 'A Lady in Distress: Inhibiting Effects of Friends and Strangers on Bystander Intervention' in *The Journal of Experimental Social Psychology* 5:189-202.
Libet, Benjamin (2002). 'Do We Have Free Will?' in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.
Mackay, Donald (1987). 'Divided Brains – Divided Minds?' in *Mindwaves*, edited by Colin Blakemore and Susan Greenfield. London: Basil Blackwell.
Manning, Rachel; Levine, Mark and Collins, Alan (2007). 'The Kitty Genovese Murder and the Social Psychology of Helping: The parable of the 38 Witnesses' in *American Psychologist* 62/6:555-562.
McKenna, Michael (2001). 'Review of *Responsibility and Control: A Theory of Moral Responsibility*' in *The Journal of Philosophy* 98/2:93-100.
McKenna, Michael (2004a). 'Compatibilism' in *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. URL = <http://plato.stanford.edu/archives/sum2004/entries/compatibilism/>. Accessed 4th August 2008.
McKenna, Michael (2004b). 'Compatibilism: State of the Art' in *The Stanford Encyclopedia of Philosophy*, edited by Edward Zalta. URL = <http://plato.stanford.edu/archives/sum2004/entries/compatibilism/supplement.html>. Accessed 10th November 2008.
Mele, Alfred. (1987). *Irrationality*. New York: Oxford University Press.
Mele, Alfred (2000). 'Reactive Attitudes, Reactivity, and Omissions' in *Philosophy and Phenomenological Research* 61/2:447-452.
Mele, Alfred (2008). 'Psychology and Free Will: A Commentary' in *Are We Free?: Psychology and Free Will*, edited by John Baer (et al.). Oxford: Oxford University Press.
Mele, Alfred (2009). *Effective Intentions: The Power of Conscious Will*. Oxford: Oxford University Press.
Nadelhoffer, Thomas (2006). 'Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality' in *Philosophical Explorations* 9/2:203-219.
Nagel, Thomas (1986). *The View From Nowhere*. New York: Oxford University Press.
Nahmias, Eddy (2001). *Free Will and the Knowledge Condition*. Doctoral dissertation, Duke University.
Nahmias, Eddy (2005). 'Agency, Authorship and Illusion' in *Consciousness and Cognition* 14:771-785.
Nahmias, Eddy (2007). 'Autonomous Agency and Social Psychology' in *Cartographies of the Mind: Philosophy and Psychology in Intersection*, edited by Massimo Marraffa (et al.). Dordrecht: Springer.
Nelkin, Dana (2005). 'Freedom, Responsibility and the Challenge of Situationism' in *Midwest Studies in Philosophy* 29/1:181-206.

Nelkin, Dana (2008). 'Responsibility and Rational Abilities: Defending an Asymmetrical View' in *Pacific Philosophical Quarterly* 89/4:497-515.
Nichols, Shaun (2008). 'How Can Psychology Contribute to the Free Will Debate?' in *Are We Free?: Psychology and Free Will*, edited by John Baer (et al.). Oxford: Oxford University Press.
Nisbett, Richard and Bellows, Nancy (1977). 'Verbal Reports About Causal Influences on Social Judgments: Private Access Versus Public Theories' in *The Journal of Personality and Social Psychology* 35/9:613-624.
Nisbett, Richard and Wilson, Timothy (1977). 'Telling More Than We Can Know: Verbal Reports on Mental Processes' in *Social Cognition: Key Readings*, edited by David Hamilton. New York: Psychology Press.
Nørretranders, Tor (1991). *The User Illusion: Cutting Consciousness Down To Size*. London: Penguin Books.
Nozick, Robert (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.
Pereboom, Derk (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
Pettit, Philip (2007). 'Neuroscience and Agent-Control' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.
Ross, Don (2007a). 'Science Catches the Will' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.
Ross, Don (2007b). 'The Economic and Evolutionary Basis of Selves' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.
Ross, Lee and Nisbett, Richard (1991). *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.
Russell, Paul (2002). 'Pessimists, Polyannas, and the New Compatibilism' in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.
Scanlon, T.M. (2002). 'Reasons and Passions' in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Skinner, Burrhus (1948/1976). *Walden Two*. Indianapolis: Hackett.
Slote, Michael (1980). 'Understanding Free Will' in *The Journal of Philosophy*, 77/3:136-151.
Smilansky, Saul. (2000). *Free Will and Illusion*. Oxford: Oxford University Press.
Sommers, Tamler (2007a). 'The Objective Attitude' in *The Philosophical Quarterly* 57/228:321-341.
Sommers, Tamler (2007b). 'The Illusion of Freedom Evolves' in *Distributed Cognition and the Will: Individual Volition and Social Context*, edited by Don Ross (et al.). Cambridge, MA: MIT Press.
Stout, Rowland (2005). *Action*. Chesham: Acumen.
Strawson, Galen (1986a). *Freedom and Belief*. Oxford: Oxford University Press.
Strawson, Galen (1986b). 'On the Inevitability of Freedom from the Compatibilist Point of View' in *Real Materialism and Other Essays*. Oxford: Oxford University Press.
Strawson, Galen (1989). 'Consciousness, Free Will, and the Unimportance of Determinism' in *Real Materialism and Other Essays*. Oxford: Oxford University Press.
Strawson, Galen (1994). 'The Impossibility of Moral Responsibility' in *Free Will*, second edition, edited by Gary Watson. Oxford: Oxford University Press.
Strawson, Galen (2000). 'The Unhelpfulness of Determinism' in *Philosophy and Phenomenological Research* 60/1:149-155.

Strawson, Galen (2002). ‘The Bounds of Freedom’ in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.
Strawson, Galen (2003). ‘Mental Ballistics: The Involuntariness of Spontaneity’ in *Real Materialism and Other Essays*. Oxford: Oxford University Press.
Strawson, Galen (2004). ‘Free Agents’ in *Real Materialism and Other Essays*. Oxford: Oxford University Press.
Strawson, Peter (1959). *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.
Strawson, Peter (1962). ‘Freedom and Resentment’ in *Free Will*, second edition, edited by Gary Watson. Oxford: Oxford University Press.
Stump, Eleonore (2002). ‘Control and Causal Determinism’ in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Suhler, Christopher and Churchland, Patricia (2009). ‘Control: Conscious and Otherwise’ in *Trends in Cognitive Sciences* 13/8:341-347.
Taylor, James (2005). ‘Introduction’ in *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy*. New York: Cambridge University Press.
Thaler, Richard and Sunstein, Cass (2008). *Nudge: Improving Decisions About Health, Wealth and Happiness*. London: Penguin Books.
van Inwagen, Peter (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
van Inwagen, Peter (2002). ‘Free Will Remains a Mystery’ in *The Oxford Handbook of Free Will*, edited by Robert Kane. Oxford: Oxford University Press.
Vargas, Manuel (2005). ‘The Trouble With Tracing’ in *Midwest Studies in Philosophy* 29/1:269-291.
Velleman, J. David (1992). ‘What Happens When Someone Acts?’ in *Perspectives on Moral Responsibility*, edited by John Fischer and Mark Ravizza. New York: Cornell University Press.
Velleman, J. David (2002). ‘Identification and Identity’ in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Wallace, R. Jay. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Waller, Bruce (1990). *Freedom Without Responsibility*. Philadelphia: Temple University Press.
Watson, Gary (1975). ‘Free Agency’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (1987a). ‘Responsibility and the Limits of Evil: Variations on a Strawsonian Theme’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (1987b). ‘Free Action and Free Will’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (1996). ‘Two Faces of Responsibility’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (1999). ‘Soft Libertarianism and Hard Compatibilism’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (2001). ‘Reasons and Responsibility’ in *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.
Watson, Gary (2002). ‘Volitional Necessities’ in *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by Sarah Buss and Lee Overton. Cambridge, MA: MIT Press.
Wegner, Daniel (2002). *The Illusion of Conscious Will*. London: Bradford Books.
Wegner, Daniel (2003). ‘The Mind’s Best Trick: How We Experience Conscious Will’ in *Trends in Cognitive Sciences* 7/2:65-69.

Wegner, Daniel (2004a). 'Précis of *The Illusion of Conscious Will*' in *Behavioral and Brain Sciences* 27/5:649-659.
Wegner, Daniel (2004b). 'Frequently Asked Questions About Conscious Will' in *Behavioral and Brain Sciences* 27/5:679-692.
Wilkes, Kathleen (1988). *Real People: Personal Identity Without Thought Experiments*. Oxford: Oxford University Press.
Wilson, Timothy (2002). *Strangers To Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.
Wolf, Susan (1980). 'Asymmetrical Freedom' in *Moral Responsibility*, edited by John Fischer. Ithaca: Cornell University Press.
Wolf, Susan (1987). 'Sanity and the Metaphysics of Responsibility' in *Free Will*, second edition, edited by Gary Watson. Oxford: Oxford University Press.
Wolf, Susan (1990). *Freedom Within Reason*. New York: Oxford University Press.
Zimbardo, Philip (2007). *The Lucifer Effect: How Good People Turn Evil*. London: Rider.
Zimbardo, Philip (2009). Interview with Tamler Sommers in *The Believer* 67. URL: <http://www.believermag.com/issues/200909/?read=interview_zimbardo>. Accessed 17th December 2009.