1

The Nature of Responsibility

The practice of holding an agent criminally responsible for breaching the criminal law is a specific instance of the more general practice of holding agents responsible for what they do. For this reason, understanding criminal responsibility requires us to understand the nature of responsibility more generally. Although the criminal justice system has an institutional character, and that distinguishes criminal responsibility from the ordinary moral practice of holding an individual responsible, the more general idea of responsibility is at the heart of criminal responsibility.

The purpose of this chapter is to develop some of the central contours of a theory of responsibility. Much of the discussion in this chapter will be to a degree preliminary and somewhat abstract. The purpose is to outline the central questions to be addressed when thinking about responsibility, questions that will be explored in more particular contexts later in the book.

Various different questions are asked in the criminal law which at least utilise the language of responsibility, and it will be important to distinguish between these individual questions. Firstly, we might ask who counts as a responsible agent as far as the criminal law is concerned. This is a question of status. For example, children below a certain age do not have the general status of being responsible agents, and that exempts them from criminal responsibility without any investigation into whether they have fulfilled the conditions of a criminal offence. Let us call this 'status-responsibility'.

Secondly, it might be important to determine the ambit of a person's responsibility. In the context of criminal omissions it is common to ask the question of whether the defendant was responsible for caring for a particular person, or for performing a particular kind of action, in order to determine whether they ought to be liable for the resultant event. So a lifeguard might be held responsible for the death of a drowning child in a swimming pool because the lifeguard is responsible for helping to ensure the safety of those in the pool. Or a parent might be held responsible for injury to his child because parents are responsible for the well-being of their children. This is commonly called 'role-responsibility'.[1] However,

[1] See particularly H L A Hart *Punishment and Responsibility* (Oxford: OUP, 1968) 212–14. I use Hart's term in 'Recklessness and the Duty to Take Care' in S Shute and A P Simester *Criminal Law Theory: Doctrines of the General Part* (Oxford: OUP, 2002).

in order to leave open the question whether this idea has scope beyond the particular roles that people play socially,[2] I will use the term 'prospective-responsibility'.[3]

Thirdly, a number of different questions are commonly asked to determine whether an event is appropriately related to the defendant such that the defendant can be held responsible for an action of a particular kind. Suppose that a death has occurred. Is the death related to the agent in such a way that it is appropriate to say that the agent is responsible for killing? The most basic element of responsibility that will help us to answer that question is that of causation, which will be the subject of Chapter 6. An agent will not normally be responsible for an event if he did not cause that event. Causation is so intimately related to responsibility that the language of responsibility is sometimes used in relation to causal relations that do not involve autonomous agents. For example, if the crack in the wall caused the building to collapse, it is common to say that the crack in the wall was responsible for the building collapsing. Where the agent has caused an event, that event can normally be re-described as an action. If the defendant has caused a death, then he has killed.

But even when a defendant has performed the kind of action that the criminal law rightly takes an interest in, the defendant may not be responsible for performing that action. For example, the defendant who was involuntarily intoxicated to the appropriate degree might not be responsible for acts performed as a result of that intoxication. Or the defendant who suffered from a particular mental disorder such that she could not appreciate the nature and quality of the act which she has performed might not be responsible for that act. This group of questions refers to what I will call 'attribution-responsibility', the conditions under which an action or event can be attributed to an agent who has the appropriate status.

Sometimes theorists of criminal law attempt to construct general theories of criminal responsibility. The most common are theories based on capacity, theories based on choice and theories based on character. Very basically, capacity theories of responsibility contend that an individual is responsible for an action only insofar as he had some or other capacity with regard to the action. This might include the capacity to have done otherwise, or the capacity to recognise the wrongfulness of what he has done. Choice theories, on the other hand, contend that an individual is responsible for his action only insofar as he chose to do it and that he has an acceptable range of choices. Finally, character theories contend that an individual is responsible for his action only insofar as his action was reflective of his character. In Chapter 2 of this book I will consider such theories. I will argue that the central ideas that motivate character theory are central to criminal responsibility, but that those ideas need to be carefully refined. However, I will also suggest that capacity has a proper place in a theory of criminal responsibility.

[2] In fact I don't think that role is the central idea here. See Chapter 7.

[3] See also R A Duff *Criminal Attempts* (Oxford: OUP, 1996) 320–2 and P Cane *Responsibility in Law and Morality* (Oxford: Hart, 2002) 31–5.

But before we can see why the broad concerns of character theorists ought to guide our account, and how they ought to be refined, we ought initially to focus on what I think is a more central and basic idea of responsibility. That idea is that an agent is responsible for an action only insofar as that action reflects in the appropriate way on the agent *qua* agent. I think that the contours of criminal responsibility that I indicated above are all underpinned by this basic claim. In order to understand the proper scope of criminal responsibility, then, we will have to know something about the nature of agency, something about what makes an action reflect on agency, and what impediments there might be to an event reflecting on the agent *qua* agent. Focusing on this basic issue will help us to determine the relevance of questions of choice, capacity and character to criminal responsibility.

In Section 1.1 of this chapter I will focus on a central case in which an agent is responsible for an action. In this central case the action of the agent can be explained in terms of the reasons that motivated him in performing his action. Such an idea, I will argue, implies that the agent's motivating reasons for action can be held up for scrutiny in the light of the normative reasons that apply to that action. This secures the centrality of intentional action to accounts of responsibility. For actions done for a reason are intentional actions.

Section 1.2 will be concerned with the challenge to this central account of responsibility from the fact that a agent may be alienated from his motivations. If an agent is alienated from his motivations in performing an action, the agent's action does not reflect on him *qua* agent. That will lead to the development of a more general account of what it means for an action to reflect on an agent *qua* agent. I will consider two well-known responses to this problem of alienation. The first is based on developing a hierarchy of desires, first proposed by Harry Frankfurt. This, I will argue, provides the basic structure of the proper response, but faces problems. These problems lead us to focus on the values of the agent, as proposed by Gary Watson. Whilst accounts based on values have been challenged in the philosophical literature, I will argue that we can refine an account based on values that meets those challenges.

In Section 1.3 I will develop two qualifications to the refined hierarchical view outlined in Section 1.2. Firstly, I will argue that there is a historical element to assessing questions of responsibility. Whether an agent's action is reflective of his agency more generally, in the way supposed by the refined hierarchical account I develop, I will suggest, depends on an investigation of the nature of the agent over time. Whether the desire under which an agent performs an action is connected to the agent *qua* agent is sensitive to the history of the desire, not merely to the values of the agent at the time at which the action is performed. Secondly, I will argue that even if the desire under which an action is performed is not related to the agent in the appropriate way, the agent may still be responsible for the action. This will be the case if the agent ought to have resisted the desire based on other normative reasons that there are against performing the action.

1.1 Responsibility and Explanation: A Central Case

When we are looking for the nature of responsibility, what kind of thing are we looking for? In some influential accounts of responsibility, the nature of responsibility is intimately connected to certain kinds of social practice, either of the agent herself, or of others. Uncovering the nature of responsibility, on these accounts, involves nothing more than developing a proper account of the social practices in which responsibility commonly manifests itself. It is true that when thinking about responsibility it makes sense not to distance ourselves too far from such social practices. However, I will argue that the basic nature of responsibility is to be distinguished conceptually from those social practices. Responsibility, I will claim, is intimately related to agency, and I will provide an explanation of what this means. This idea can fruitfully be developed in relation to social practices, but those social practices can also best be understood once we understand that their proper object is agency.

There is an obvious etymological connection between responsibility and the idea of a response. It is that connection that leads some to consider different kinds of social practice that are said to reveal something further about the nature of responsibility. Two different kinds of response have been considered relevant in unravelling the nature of responsibility. First, there is the response that an agent might be required to make in relation to some action that he has performed. Using this idea in a definitional way, we might claim that to be responsible is intrinsically connected to the practice of giving an account of oneself. For example, Antony Duff writes that 'to be responsible is to be answerable'.[4] To ask who is answerable, under what circumstances, for what events, is, on this account, to ask who is responsible, under what circumstances, for what events.

Second, we might focus not so much on the response of the agent to his or her conduct, but rather on the responses that others might give to one's conduct. This account has become familiar through the well-known work of P. F. Strawson.[5] The relevant responses, on Strawson's account, are both emotional and social. To be responsible, Strawson claims, is to be an appropriate target for 'reactive attitudes'. Reactive attitudes include attitudes such as resentment and condemnation as well as gratitude or approval. And these reactive attitudes generate the social practices of praising or blaming the individual.[6] This account is developed as a response to the supposed challenge to responsibility posed by determinism. Strawson attempts

[4] R A Duff *Punishment, Communication, and Community* (Oxford: OUP, 2001) 184.

[5] See 'Freedom and Resentment' in G Watson *Free Will*, 2nd edn. (Oxford: OUP, 2003) and also the book-length development of Strawson's view R Jay Wallace *Responsibility and the Moral Sentiments* (Cambridge, Mass.: Harvard University Press, 1996).

[6] We might add that there are reactive attitudes and practices of the agent herself. The agent might feel regret or pride for those things for which she is responsible, and might punish herself or reward herself accordingly. See B Williams *Shame and Necessity* (Berkeley: University of California Press, 1993) 55.

to show that the social practices that constitute responsibility suppose nothing about determinism. In fact, he claims not to know what determinism is.[7]

Both of these ideas seem central to a full account of responsibility, and have furthered understanding of responsibility enormously. A proper theory of responsibility should help to illuminate the meaning of such reactions, both emotional and social, of others. It should help us to see why those reactions have responsibility as their object. Furthermore, both these accounts can accommodate something that is of central importance about responsibility: the idea of responsibility does not in itself fix the *appropriate* kind of response. To say that an individual is responsible for an action is not yet to say what kind of explanation is appropriate. It does not in itself determine whether that individual is *at fault* for what he has done, or whether he is to be exonerated. Hence, those who are justified or excused are often responsible for their actions: justification and excuse provide at least part of the explanation that responsible agents must provide if they are to avoid blame for what they have done.[8] Consequently, that an agent is responsible for an action does not, in and of itself, make any particular reactive attitude appropriate. The idea of responsibility leaves open whether we should feel resentful toward the agent on the one hand, or grateful on the other. It leaves open the possibility that it is appropriate to feel sympathy for the agent or indignant about his failings. That an agent is responsible for an action only puts us on notice that some kind of social and emotional reaction may be appropriate with regard to the agent for performing a particular action.

Despite the intimacy of the relationship between responsibility and our social and emotional practices, however, those practices invite us to provide a further account of something more basic: *what it is* that we are responsible for. For they invite us to ask *what it is* that we can properly be called to account for, or that we react to in the significant sense. I do not wish to suggest that either theory of responsibility is especially problematic in itself, although as we shall see, there may be exceptions where the agent is responsible but is not required to answer, and where the agent is responsible but we do not make the relevant emotional response to his conduct. But each theory invites us to explore something more basic about the nature of responsibility.

Let us reflect a little further on the thesis that responsibility is fundamentally connected to answerability. The agent, it is claimed, is responsible for those things that he must answer for. But we are entitled to ask *what it is* that the agent must answer for. We cannot focus on the nature of the social practice without thinking about its object. Hence, the nature of the social practice cannot be used to generate the nature of responsibility. On the contrary, it is the nature of responsibility that generates a full theory of answerability.

[7] I will have more to say about what determinism is in Chapter 2.

[8] See also T M Scanlon *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998) 248 and J Gardner 'The Mark of Responsibility' (2003) 23, *Oxford Journal of Legal Studies* 157.

Similarly, in relation to the thesis that responsibility is fundamentally connected to the reactive attitudes, surely we are entitled to ask *what it is* that is the appropriate target for the reactive attitudes and the practices that they generate. It is not circular to answer 'the actions, beliefs and emotions for which the agent is responsible'. For it is only that idea that distinguishes between the *proper* object of both answerability and the reactive attitudes, and improper objects of those practices. We can only focus on the reactive attitudes once we know what they react to. This suggests that we will need to look more deeply at the *appropriate object* of the social practices of providing an account of oneself, or the reactive attitudes and the social practices of praise and blame that they generate, in order to discover the nature of responsibility.

We can reveal why it is appropriate to begin with the basic question of the nature of responsibility even more clearly by considering the fact that it is possible to hold apart our social practice of accounting and our emotional reaction to behaviour on the one hand, and the idea of responsibility on the other, at least with regard to particular cases. Even if we fail to call a particular agent to account for an action, or if we take a more 'objective stance', distancing ourselves from the ordinary emotional reaction to certain kinds of behaviour, that is not to say that we immediately lose the sense that the agent is responsible for that action. The idea of responsibility appears to be sufficiently robust that it can survive the absence of the kinds of social and emotional practices in which it normally becomes manifest. Of course, that is not to say that the existence of emotional and social practices *in general* does not play a central role in unveiling the idea of responsibility. The more basic account to be given of responsibility will not be fully explicable without presupposing at least the possibility of answerability on the one hand and the reactive attitudes of others on the other.[9]

Finally, there are emotional attitudes that *are* appropriate with regard to behaviour which the agent is not responsible for.[10] And sometimes an agent may have to answer for behaviour for which he is not responsible. To take reactive attitudes first, it may be that we respond emotionally to an event which befalls the agent by thinking that the agent is lucky or unlucky, and this may prompt an emotional reaction. We may feel sympathy for him when he is unlucky, or we may feel envy towards him when he is lucky. The relationship between luck and responsibility is complex,[11] but there are at least some instances in which those attitudes are appropriate where the agent is not responsible for the relevant event.

[9] See G Watson 'Responsibility and the Limits of Evil' in J M Fischer and M Ravizza *Perspectives on Moral Responsibility* (Ithaca: Cornell University Press, 1993).

[10] Indeed, Strawson himself notes that there are emotional responses that are appropriate for those who lack status-responsibility, as well as the actions for which those with status-responsibility are not responsible. See 'Freedom and Resentment', 77–80.

[11] See, for example, T Nagel 'Moral Luck' in *Mortal Questions* (Cambridge: CUP, 1979), B Williams 'Moral Luck' in *Moral Luck* (Cambridge: CUP, 1981) and S Hurley *Justice, Luck, and Knowledge* (Cambridge, Mass.: Harvard University Press, 2003). I consider this question in the context of the criminal law in Chapter 3.

Undoubtedly, these reactive attitudes are to be distinguished from attitudes such as resentment or gratitude. However, in order to develop an account of responsibility, we must discover what the proper basis of that distinction is, which suggests something deeper and more basic than the mere existence of the emotional responses, and the social practices that go with them.

Turning to the requirement to answer, it may be that there are circumstances in which a person is required to give an explanation for conduct where it turns out that he is not responsible for that conduct. Consider the agent who attacks another under the influence of a very powerful hallucinogenic drug that has been surreptitiously administered to him. That agent is not responsible for his conduct. And yet that is not to say that it is inappropriate to call him to give an account for that conduct. He is properly required to give an explanation for the attack, but if the explanation is of the appropriate kind, we do not hold him responsible for the conduct. That is not to deny that there is some sense in saying about this agent that he is not 'answerable' for his conduct, but we need to know something further about the *kind* of answer that is being asked for in accounts of responsibility.

To summarise, a basic account of responsibility ought to provide an explanation of different ways in which we might react to the agent, and different ways in which he is to be invited to account for his conduct. That is not to say that these phenomena are not central to developing a proper account of responsibility. These social and emotional practices can help us to make progress in illuminating the more basic issue of responsibility. However, we need to know something further about what kind of account will be provided, and what kind of target the reactive attitudes have, in cases where the agent is responsible for his conduct. This will help us to unravel the nature of responsibility.

Let us reflect a little more on the nature of the response that a person typically gives when he is held responsible. The relevant response is normally to a demand for a particular kind of explanation. Where I am responsible for a particular event, and an explanation is demanded of me, one common response is to give that explanation in terms of my motivations: that is, to the reasons why, at least at the time of action, I thought that the action I performed was worth performing.

Such an explanation is normally provided where my action was intentional. Intentional action is commonly at the centre of theories of responsibility, and if a theory of responsibility cannot accommodate responsibility for intentional action it would be seriously deficient. My intentional actions are to be distinguished from actions that are not performed intentionally by virtue of the fact that intentional actions can be explained in terms of the reasons that I had for performing them.[12] That is not to say that I *want* to perform all of my intentional actions, in the sense that I do them without reluctance, but I can at least provide an explanation for them in terms of my motivations which make clear why, at the time of action, I thought that there was something positive to be said for performing the action.

[12] See, further, Chapter 7.

From this central case, we can begin to illuminate our account of responsibility further: the explanation of an event that is required in cases of responsibility is often explanation in terms of at least some *psychological* features of the agent. In referring to the agent's psychology, I do not mean any deep level and hidden desires of the agent, but rather attributes of the mind of the agent. In fact, the relevant kind of explanation is precisely to be distinguished from explanation in terms of the agent's 'unconscious self', if indeed there is such a thing. When the agent gives his reasons for action, he is normally talking about an action that is guided at the level of consciousness,[13] albeit not necessarily with a very high degree of reflection.[14]

Given this, we can see some reasons to be careful about the terminology that we use with regard to reasons that explain action in the relevant sense.[15] In the literature on intentional actions, the term 'explanatory reasons' is sometimes used to refer to the reasons that operated in producing the action at the level of consciousness. There is something appropriate about this terminology. When an agent acts for a reason, that reason provides one way in which the action can be explained. But we should not suppose that that is the only potentially relevant kind of explanation of an action. For example, we might explain D's *v*ing at the level of neuroscience, or in terms of the agent's unconscious. Such explanations do not necessarily suppose that explanation in terms of the agent's conscious psychology is inappropriate. However, those kinds of explanation are to be distinguished from the kind of explanation that refers to the reasons that D recognised at the time of action as making *v*ing worthwhile.

In referring to motivating reasons, we are looking for *rational* explanation as distinct from scientific or deep psychological explanation, where rational explanation involves psychological states at the level of consciousness. Rational explanation appeals to the agent's ability to evaluate his action. As Philip Pettit puts it:

> Rational explanation of action involves the attempt to explain an agent's speech or behaviour by reference to distinctive psychological states: roughly, by reference to states that reflect the information to which the agent gives countenance and the inclination that moves him or her; by reference, as the stock phrase has it, to beliefs and desires.[16]

For this reason I prefer to use the term 'motivating reasons' rather than 'explanatory reasons'[17] in this context. The former term implies something about the kind of explanation that is at issue.[18]

[13] In fact, the agent normally has a particularly intimate knowledge of his intentions. A person is not normally uncertain about his intentions. It does not generally make sense to say 'I believe that I have the intention to *v*, but I am not sure.' See L Wittgenstein *Philosophical Investigations* trans. G E M Anscombe (Oxford: Blackwell, 1953) 247.

[14] As I will show in Chapter 8, intentional action itself often does not require a high level of reflection on the part of the agent.

[15] See also J Dancy *Practical Reality* (Oxford: OUP, 2000) 6–7.

[16] 'Three Aspects of Rational Explanation' in *Rules, Reasons, and Norms* (Oxford: OUP, 2002).

[17] For example, see J Gardner 'Justifications and Reasons' in A P Simester and A T H Smith *Harm and Culpability* (Oxford: OUP, 1996).

[18] Of course, even that term is not perfect. We may say that fear was the reason D jumped out of the car, or that D was motivated by fear in jumping out of the car. But that is not the kind of explanation that we ought to be looking for either.

When we hold an agent responsible for an action, then, we suppose that the action can be explained in terms of the psychology of that agent. The central case of such explanation is by reference to the motivating reason for which he acted. This can help to illuminate the idea of accountability. When the agent is held to account for his action, he is required to provide an explanation. That explanation is normally in terms of features of his psychology, and in particular the reasons that motivated the action.

But when we hold an agent accountable for his action, we do not *merely* demand an explanation of his action. The agent's explanation makes sense only because we think that there are norms that apply to the agent. Ideally, at least, the agent will be able to show that the reasons that motivated her corresponded to the norms that applied to her at the time of acting. She shows that the psychological explanation of her action corresponded to the normative demands of the circumstances that she was in. In other words, she is asked to show that her motivating reasons corresponded to normative reasons: reasons that ought to have guided her action.[19] That an agent was motivated to act in a particular way suggests that she thought that there were normative reasons for action: that, in some way or other, the action was worth performing. Of course, she need not have believed that the action was particularly *moral*. But she must have seen something about the action that made its performance attractive. Providing such an explanation supposes that we can establish whether she was right about that. In short, we hold the motivating reasons of the agent up to scrutiny in the light of *normative* reasons.

This relationship between motivating reasons and normative reasons can help us to understand something further about the reactive attitudes of agents. Let us focus on negative attitudes such as resentment in the context of action. Broadly speaking, attitudes such as resentment towards an agent are appropriate when the normative demands that applied to the agent in the circumstances in which he found himself do not correspond to the way in which that agent was motivated in performing his action. From this, we can see that resentment is properly directed at the agent in virtue of the connection between the action and his psychology. We do not resent an agent simply for a result, we resent him for performing an action: an event that can be explained in terms of features of the agent *qua* agent.[20]

At least one central case of responsibility for an action, then, is that of intentional action. That is, action performed for a motivating reason. When holding an agent responsible for an action, in this central case, we scrutinise the reasons which motivated his action by holding them against normative reasons. And that normally demands that the agent provides an account of his action in terms of his motivating reasons. Ideally, at least in the case of the criminal law, holding the agent responsible will also involve communication with that agent about the

[19] See also J Gardner 'Justifications and Reasons'. Gardner uses the term 'guiding reason' rather than 'normative reason'. Although we disagree about the appropriate terminology and what constitutes a normative reason, Gardner's account has influenced mine.

[20] Beyond the context of action, we might resent an agent simply for the attitudes or beliefs that he has.

appropriate norms, and whether his intentional actions were justified in the light of those norms. Where, following such communication, normative and motivating reasons come apart, it is often appropriate to react to the agent negatively, say with resentment or indignation. Such reactive attitudes often appropriately have as their target the performance of an action under the guidance of the agent's psychology in the sense outlined.

So far we have focused on intentional action in our account of responsibility. But responsibility is not limited to intentional action. For one thing, accounts in terms of responsibility involve things that are not done intentionally. Lawyers will be familiar with a range of actions beyond intentional actions, but which are commonly the subject of legal responsibility of some kind. First, there are events that the agent does not intend, but knows will come about as a result of his intentional actions. These are commonly, though as we will see in Chapter 8 not ideally, called actions done with 'foresight'. Second, there are actions which the agent risks doing, commonly called cases of 'subjective recklessness' by lawyers. Thirdly, there are actions which the agent does without realising she would do them, but where she ought to have realised she would do them, commonly called cases of 'objective recklessness', or 'negligence'. Finally, there are actions which it matters not the state of mind with which they are done. This is true for offences of absolute liability, where responsibility can arise without reference to the agent's psychology. I will not consider this latter category in detail in this book. Much of this book will be devoted to considering whether criminal responsibility is appropriate in these different cases and I will say little further about these actions here.

Beyond action, we ought to consider responsibility for other kinds of event or state, particularly for the features of the agent's psychology themselves. An agent may be held responsible not only for his actions, but also for his desires, his beliefs and his emotions. Our basic account of responsibility can help us to illuminate why this is so. Just as action can be explained by motivating reasons, so can desires, beliefs and emotions. It is common to ask of an agent why he wanted what he wanted, why he believed what he believed or why he felt what he felt. In that case, we look for an explanation of that desire, belief or emotion. And that explanation will normally be in terms of the motivating reasons of the agent.

Furthermore, for such features of the agent's psychology, it is appropriate to hold the agent's account up to scrutiny in relation to normative reasons. If the agent suggests that he wanted to *v* because of *r*, it is appropriate to ask whether *r* constituted a good reason to want to *v*. If the agent suggests that he believed *p* because of *r*, it is appropriate to ask whether *r* constituted a good reason to believe that *p*. If the agent suggests that he felt *e* because of *r*, it is appropriate to ask whether *r* constituted a good reason to feel *e*. We may, at least in some circumstances, appropriately call the agent to account for his desires, beliefs and emotions as well as for his actions. He can be expected to provide an explanation for those desires, beliefs and emotions in terms of motivating reasons. And we can scrutinise those reasons in the light of normative reasons.

Finally, reactive attitudes may be appropriate with regard to desires, beliefs and emotions as well as with regard to actions. We may feel that the agent is superficial for wanting money more than love. We may feel that she is gullible for believing that her husband was faithful because he told her so. We may feel that she is short-tempered because she went into a furious rage at another driver for failing to pull away quickly enough at the lights. And we may blame her for all these reactions. Our basic account of responsibility, then, is as applicable with regard to desires, beliefs and emotions as it is with regard to actions. In each case, it is often appropriate to scrutinise the motivating reasons of the agent in the light of normative reasons, to demand an explanation and to react both emotionally and socially in the light of that explanation.

1.2 Refining a Hierarchical Account of Responsibility: Responsibility and Value

An agent is responsible for an action, belief, desire or emotion, I have claimed, insofar as that action, belief, desire or emotion reflects on the agent *qua* agent. There are a number of different ways in which events or states of affairs may reflect on me. For example, the extent to which I am tall or lucky reflects on me in some way. But they do not reflect on me *qua* agent. For this reason, they are also inappropriate targets of *normative* scrutiny. That is not to say that we should not value being tall or being lucky, or that it is always inappropriate to react emotionally and socially to height or fortune. But we should not confuse those reactions with the kinds of reactions that are appropriate when agency is at issue.

In focusing on actions, beliefs, desires and emotions, we plausibly focus on the constituents of agency. And that is shown by the fact that one acts, believes, wants and feels for reasons. Actions, beliefs, desires and emotions, I have claimed, can be explained in terms of the reasons that motivate the agent, and can be scrutinised in the light of normative reasons. It is this connection that makes actions, beliefs, desires and emotions appropriate subjects for praise or criticism in themselves.

Why should explanation in terms of motivating reasons ground the idea of responsibility? The obvious answer is that motivating reasons are constituents of agency. Insofar as an action is performed under the guidance of a motivating reason of the agent, it might be thought, that action is performed under the guidance of the agent. And that grounds the agent's responsibility for the action.

However, this account is insufficient in itself. For although it is true that motivating reasons are often constituents of agency, it may be that the agent acts for a motivating reason, but that motivating reason is not appropriately connected to the agent *qua* agent. The agent may be alienated from the reason that motivated the action. Much of the rest of this chapter will be concerned with understanding how an agent can be alienated from his motivations, and the significance of alienation for an account of responsibility.

Firstly, let us illuminate the issue with the following example. D has his drink spiked with a drug that creates a very powerful desire on his part to *v*. He has no control over the desire. Although he does not value the desire, and would rather be rid of it, he can do nothing to remove it. He comes across an opportunity to *v*, one where there are no strong reasons against *v*ing.[21] He *v*s. In this case D is not responsible for *v*ing. In such cases, although the agent can provide an explanation of his conduct, beliefs and desires in terms of motivation, those motivations are not *his own*. If the responsibility of an agent for a particular action is grounded in the extent to which that action reflects on the agent *qua* agent, in this case there is good reason to suppose that the agent is not responsible for his action. Whilst the action reflects on the existence of the desire, the desire is not properly reflective of the agent *qua* agent.

There is even a question about whether it can *really* be said that he acts for a motivating reason at all. To act for a reason, I suggested, involves recognising at least that there is something worthwhile in performing the action. Here the agent acts under the guidance of a desire. But desires do not generally provide reasons for action. It may be that the agent has a desire, but has no reason to fulfil it. Nevertheless, the agent can at least provide an explanation of the action in terms of his psychology. And that explanation does not ground his responsibility. For the desire which guided the action was disconnected from him *qua* agent. I will use the term 'motivating reason' in a broad sense, which covers such cases.

This suggests that those who defend the character theory of criminal responsibility have at least something right. When thinking about whether the agent is responsible for his action we need to look more deeply than whether the action was performed for a motivating reason, we need to know whether that motivating reason reflected on the agent *qua* agent.[22] It remains to be seen whether the concept of 'character' does the appropriate work to capture this idea.

A common response to this problem in the philosophical literature is to develop a hierarchy of motivation. In thinking about agency, we need to reflect not only on the agent's motivating reasons, we need to think about his attitude towards those motivating reasons. And this, it is argued, can secure the appropriate relationship between the agent and his motivations to ground responsibility. The classic statement of this thesis was developed by Harry Frankfurt.[23]

Suppose that an agent, D, *v*s, and his *v*ing is motivated by a desire *d*. In that case, *d* will provide the explanation of the agent's *v*ing. But, as we have seen, that may not make D responsible for *v*ing. For he may be alienated from *d*. How are we to distinguish between desires from which the agent is alienated and desires which

[21] This must be stipulated, as if there was a strong reason against *v*ing we might expect D not to *v* if he was properly motivated. In that case, *v*ing would reflect on D. His desire to *v* would not reflect on him, but that he *acted* on the desire would. See below for further discussion. The relevance of this idea to the law will be investigated in Chapter 12.

[22] See also R A Duff *Criminal Attempts* 189. I will consider character theories of criminal responsibility in more detail in Chapter 2.

[23] 'Freedom of the Will and the Concept of a Person' in *The Importance of What we Care About* (Cambridge: CUP, 1988).

he is not alienated from? Frankfurt's answer is to introduce a further level of desire. In order to be responsible for *v*ing, in such a case, D must not only have the desire to *v*, he must desire that he desire to *v*, and desire that that desire is executed in action. The desire to desire to *v*, coupled with the desire that it be executed in action, Frankfurt calls a 'second-order volition'. What distinguishes responsible agents from non-responsible agents, Frankfurt thinks, is that they have the capacity not only to act on their desires, but to affirm or deny their desires, and their execution in action, in the light of higher-order desires that they have. For Frankfurt, the formation of second-order volitions distinguishes full persons from other agents, and secures the freedom of the will of those persons.

There are a number of distinct problems with Frankfurt's account which require a solution to move us towards a more acceptable and complete account of responsibility. I will consider three of those problems here. The first, which is related to the argument sometimes raised by those who think that moral responsibility and determinism are incompatible,[24] concerns the possibility of infinite regress. Suppose that the agent has a second-order volition regarding his first-order desire. In order for that second-order volition to be freely held, surely it needs to be affirmed by a third-order volition. This can be brought to light by the possibility that the second-order volition may have been detached from the agent *qua* agent in much the same way as a first-order desire. For example, suppose that the agent is hypnotised into desiring his desire to *v*, and that the desire be executed in action. He now has both a desire to *v* and a desire to desire to *v*, and to execute that desire. However, if he *v*s in the circumstances above, surely he does not *v* freely. But to require a third-order desire to affirm his second-order desire would obviously have regressive implications. Eventually, the agent will run out of desires, and consequently, if Frankfurt's view is consistently to be maintained, there appears no obvious foundation for his responsibility either. Frankfurt attempts to answer this worry by suggesting that the second-order volition must be *decisively* in favour of the desire to *v*.[25] But why should the agent not be alienated from this decisiveness itself?

A second problem with Frankfurt's account is that it overly restricts responsibility. Suppose that there is an agent, D, who desires to *v*. He would rather not have the desire to *v*. But not desiring to *v* would take some effort on D's part. His desire not to have the desire to *v* is insufficiently powerful to motivate him to remove the desire to *v*. In that case, surely D is responsible for *v*ing. Frankfurt thinks that an agent who acts on such a desire does not act freely[26] and his account implies that an agent cannot be fully responsible for an action where his desires do not converge. But as I suggested in my account of responsibility, an agent's desires may be held autonomously or they may not be. If an agent simply accepts that his desires do not conform with each other, surely that is sufficient to regard them as

[24] See, for example, G Strawson *Freedom and Belief* (Oxford: OUP, 1986).
[25] 'Freedom of the Will and the Concept of a Person' 21.
[26] 'Freedom of the Will and the Concept of a Person' 18–21.

autonomously held. It is not necessary that the agent must in fact have *achieved* conformity in his desires in order to make him responsible for them. At present, it is sufficient to note that scrutiny and regulation of one's desires is one of the key ways in which an agent can show virtue. A failure properly to regulate one's desires may correspondingly constitute a vice.[27]

A third difficulty for Frankfurt is that his account of responsibility does not involve a historical dimension. If free action is to be connected to agency, the desires that an agent has must be appropriately connected to the agent *qua* agent. But whether this is the case must in part depend upon the history of the desire. The identity of an agent persists over time. So whether a desire is reflective of the agent *qua* agent will surely be sensitive to this feature of agency. We can see some of the potential implications of this idea if we consider cases of personality change. Consider an agent who, at time *t*, does not desire to *v*. At *t2* he receives a blow to the head. Immediately after, at *t3*, he has a powerful desire to *v*, and he *v*s. It may be that D, at *t3*, is identical in all respects to D2, who performs an identical action. But the history of the way in which D came to desire to *v* and act on that desire surely undermines his responsibility. It is not really him, we might say.

I will consider the contours of the third difficulty in the next section. For now, let us focus on the first two problems. The solution that we find to the first problem will help us to see how best to resolve the second. One method by which we might attempt to solve the problem of regress, which has been suggested by Gary Watson,[28] is to distinguish qualitatively between different levels in the hierarchical order. We might distinguish between desires, which operate at the first order, and *values*, which operate at the second. An agent is responsible, on this account, if he acts under a desire to *v*, and values that desire.[29]

Now, in relation to the first difficulty that I suggested Frankfurt faced, Watson's proposal is clearly an improvement. Recall that one reason to develop hierarchical theories of responsibility is the possibility that an agent will act according to his desires, but that those desires will not be reflective of the agent *qua* agent. The agent's desire, the suggestion goes, may be alienated from the agent. Consequently, we need to establish something more about the relationship between the agent and his desire to secure the agent's responsibility for acting under that desire. However, as the first objection shows, Frankfurt's solution is insufficient to secure the appropriate relationship between the agent and his desire. For the agent may be alienated from his second-order volitions just as he is alienated from his first-order desires.

In focusing on values rather than second-order desires, Watson makes a significant advance: for the values of the agent appear to be more central to the nature of

[27] The idea that desire is a proper object of virtue owes much to Aristotle. See particularly *Nicomachean Ethics*, trans. T Irwin (Indianapolis: Hackett, 1985) book 7.

[28] See 'Free Agency' in *Free Will*, 2nd edn.

[29] Perhaps this distorts Watson's account a little, for, at 340, Watson claims that to value something *is* to desire it. The distortion, if it is such, is not significant for the account developed here.

the agent *qua* agent than his desires. Hence, whilst Frankfurt faces the problem of the agent who is alienated from his desires, Watson might not face a parallel problem of the agent who is alienated from his values. For having a value, it might be said, is constitutive of the agent *qua* agent in a way that is not true of having a second-order volition.[30] Valuing something is more intimately related to agency than desiring something.

One reason why we might think that this is so has to do with the nature of valuation. When we evaluate we consider the worth of something. And considering the worth of something involves holding that thing up to scrutiny in the light of other things that we value. Hence, valuation requires one to have a *system* of values. It may be that if one's putative valuing of one thing is not interwoven into that system, it does not count as valuing at all. However, let us leave open for a moment the possibility that an agent can value *v*ing, but be alienated from his valuing of *v*ing. For a complete response to this problem can only be appreciated in the light of our answer to a further problem with focusing on the relationship between values and desires which will have to be negotiated if the account is to stand up as an account of responsibility.

This problem concerns the second objection that I raised in relation to Frankfurt's model. In that objection, I suggested that we cannot tie our account of responsibility to actual conformity in the hierarchy of desire. The reason for this is that an agent may be responsible for making his desires conform with each other. If the agent merely accepts such lack of conformity in his desires, his responsibility for the action will not have been undermined. This problem is exacerbated once we introduce the idea of second-order valuation. The reason for this is that coherence between desires is not the primary way in which we ought to think of an agent's ethical relationship with his desires. Rather, the important relationship is between desires and normative reasons. In response to the question 'why do you want to *v*?', the agent may refer to another desire: 'because I want to *w* and I can *w* by *v*ing'. But commonly the agent is being asked more than merely to show coherence in his desires, he is being asked to show that there is something valuable about his desires. As I noted above, desires, just like beliefs and actions, can be scrutinised in the light of normative reasons.

We can now see the difficulty with Watson's account. Consider the agent who has appropriate values, but whose desires do not conform to his values. Such an agent might still be responsible for actions performed under the guidance of his desires despite this, for he might simply accept the lack of conformity between his desires and his values. If moral responsibility is to rest on the correspondence between the desire which motivated the action and the agent's values, an agent will not be responsible for any failures to make his desires conform to his values, and the actions which ensue. And that fails to recognise a central feature of ethical life: the requirement that one must scrutinise and adjust one's desires in the light of normative reasons.

[30] Watson contrasts his view with Frankfurt's in 'Free Agency' 348–50.

It is worth noting here that the failure properly to recognise this idea also leads Watson to collapse cases where the agent is compelled to have a particular desire and cases where the agent has a desire about which he is merely weak. Watson recognises that there is an important distinction between compulsive action and weak action.[31] But he does not draw the same contrast at the level of *having* the desire in the first place. It is true that, in each case, the agent's desires and evaluations fail to conform to each other. Both the weak and the compelled fail to value the desires that they have. Given this similarity, Watson argues only that there is a difference between the *action* of the weak and the *action* of the compelled. Compulsive action, he thinks, is the action of an agent whose desire does not correspond to his values, and which is so strong that even those with all of the appropriate virtues of self-control would resist them. This is to be contrasted with weak action. A weak action, for Watson, is an action that is performed on the basis of a desire that does not conform to the values of the agent, but which a virtuous agent would resist performing. It is, of course, true that an agent whose desire is resistant to her evaluations is responsible for that action if the virtuous agent would have resisted such an action. I will consider this issue in the next section. Here I will show that we can also find an appropriate way of distinguishing between weak desires and compulsive desires.

One natural way to characterise the distinction between weak and compulsive desires is to distinguish between desires that the agent could have controlled had he wished to and those that he could not have controlled.[32] This approach ultimately fails, but the failure will help us to develop a better account. The reason why the approach fails is that, in determining questions of responsibility, we ought only to be interested in the actual relationship between the agent's psychology and her desire, not her capacity to have done otherwise. This idea will be of significance when we come to evaluate choice theories of responsibility in the next chapter.

We can see the problem clearly from the following example. Consider two agents, Nora and Dora, each of whom have the desire to *v*. Nora and Dora do not value their desire to *v*. Both Nora and Dora think that they could remove their desire to *v* with sufficient effort. However, neither Nora nor Dora can be bothered to do so. The difference between them is that, unbeknownst to Nora, if she attempted to remove her desire to *v* she could not do so. Dora, on the other hand, could do so if she so wished. Both Nora and Dora act under their desire to *v*.[33]

[31] In his 'Skepticism About Weakness of Will' (1977) 86, *Philosophical Review* 316.

[32] This is the approach taken in J Kennett *Agency and Responsibility: A Common-Sense Moral Psychology* (Oxford: OUP, 2001) 159–69. Her account is motivated by the same kind of objection to Watson as mine, but her account is susceptible to the problems raised in Frankfurt cases. A similar objection arises in respect of M Smith 'A Theory of Freedom and Responsibility' in G Cullity and B Gaut *Ethics and Practical Reason* (Oxford: OUP, 1997).

[33] Readers familiar with the literature on free will and responsibility will recognise this as a particular variation on what are called 'Frankfurt cases'. I will discuss such cases in more detail in the next chapter.

There is good reason to suppose that both Nora and Dora are responsible for *v*ing. The fact that, unbeknownst to Nora, she could not remove the desire to *v*, cannot have significant impact on her responsibility, as she has not attempted to exploit that option. Nora is not truly compelled to have the desire that she has, even though she could not remove that desire even if she were minded to. Consequently, her action performed under the control of that desire is not compelled either. So an agent will be responsible for an act under a compelled desire if she in fact accepts having that desire, even if, were she not to accept having that desire, she could do nothing about it.[34]

Compulsive desires, then, are those that the agent does not accept. In this, they are opposed to cases of weak desire, which the agent accepts, even if it does not conform to his evaluations. As both Nora and Dora accept their desires, they are examples of weak desires even though only Dora would have been able to abandon her desire had she wanted to. Even if, at the moment of action, the agent's desires were sufficiently strong that they could not be resisted, if the agent simply accepts the failure of her desires to correspond to her evaluations, neither her freedom nor her responsibility is undermined. Hence, on my account, in contrast to Watson's, there is a distinction to be drawn between the desires of the weak and the desires of the compelled, as well as between the actions of those individuals. Compelled desires are desires that the agent does not accept, but reasonably fails to remove. Weak desires are desires that the agent accepts, even if they do not correspond with her values.

The account of responsibility that I am proposing suggests the following, then. If an agent performs an action under the guidance of a desire, he will normally[35] be responsible for that action unless he is alienated from that desire. An agent is alienated from a desire insofar as that desire is accepted by the agent, even if the desire does not correspond to his values. This account accommodates responsibility for agents whose desires do not conform to their values, if the agent is merely weak in accepting such conformity. And this corresponds to the fact that we are responsible for our desires: for regulating them in the light of normative reasons. This account is hierarchical. Its basic structure is similar to that proposed by Frankfurt and Watson, but I have refined that account.

Now, let us return to the problem that I raised earlier, the problem of the individual who is alienated from one of his values. Some writers suppose that it is possible to conceive of an agent who is alienated from his values, just as it is possible to conceive of an agent who is alienated from his second-order volitions. For example, David Velleman has suggested that an individual who 'recoils from his own materialism or his own sense of sin' is an example of this.[36] But if we take seriously the second objection that I raised with regard to Frankfurt's model, we can see that

[34] See also the discussion in S Hurley *Justice, Luck, and Knowledge* ch. 2.

[35] See the qualifications that I develop below to see why this is only normally the case.

[36] 'What Happens When Someone Acts?' in *The Possibility of Practical Reason* (Oxford: OUP, 2000) 134.

there is some reason to suppose that these are not true examples of alienation. For in these cases, whilst the agent may recoil from his materialism or his sense of sin, that is not to say that he does not accept it. It is only if he attempts to become less materialistic and reasonably fails that we can truly say, in such a case, that the individual is alienated from his materialism.

Earlier I talked about the systematic quality of valuing. In developing one's system of values one aims at coherence. But that is not to say that all of one's values need to achieve coherence in this way to ground responsibility for actions guided by those values. That a person is inconsistent in his system of values is often the source of criticism of that person. There may be agents who have an inconsistent set of values which they are aware of. Such a person may recoil from valuing something that they value, but that is not to say that they lack responsibility for regarding that thing as valuable. A person may place too much value on material goods and yet suspect at some psychological level that this is inconsistent with other values that they hold. But that does not rule out responsibility for acts guided by their valuing of material goods.

But now, if it can be imagined, consider the agent who becomes materialistic in some way for which he is not responsible,[37] who makes a reasonable attempt to become less materialistic, but who cannot alter his materialism. Perhaps there may be cases of brainwashing that are like this. Could we not then say that the agent is alienated from his values? Two responses to this problem may present themselves, each of which ought to be rejected. The first is to look for some further level of valuation. That must be rejected, for it would lead us straight back into the familiar problem of regress. How many levels of valuation do we need in order that it is no longer possible that the agent is alienated from his valuation?

Another option is to look for some even more basic feature of agency. It is at least plausible to hold open the possibility of an agent being alienated from one of the values that he holds. So perhaps we should try to discover something that is so intimately connected with the agent *qua* agent that there is no possibility of its being alienated from agency. This leads Velleman to attempt to discover a mental state that is incapable of being scrutinised. For, he thinks, if an element of agency is capable of being scrutinised, the agent may also be alienated from it. He posits for this role the concern to act in accordance with reasons. This concern, Velleman thinks, drives all practical thought, but we cannot be alienated from it.[38] However, it seems rather difficult to give any real flesh to this idea without recourse to the values that the agent in fact adopts. The nature of the agent who commits to Velleman's idea without committing to any values is quite mysterious. An agent who is committed to acting in the light of right reason, but does not have any of the values which would give content to this idea, cannot be imagined. Such a mystery seems to provide insecure foundations for a theory of responsibility. We do better, I think, to seek a foundation for responsible agency that is not independent from valuation in this way.

[37] I discuss this condition in Section 3 of this chapter.
[38] 'What Happens When Someone Acts?'

In order to make some progress, let us return to the account of valuation that I proposed earlier. There, I suggested that the reason why it is difficult to imagine an agent being alienated from his valuing has to do with the nature of evaluation. When an individual values something, that normally supposes that he has reflected on the worth of that thing, and in relation to other things. Valuing tends to be systematic in a way that desiring is not. That is not to say that all values are generated from a single value, or that values logically entail one another. But valuation tends to involve coherence.[39]

Now, it might be argued that the idea of systems of valuation cannot rescue the account from the challenge of regress. For even if one could show that one cannot be alienated from a single value, surely one could be alienated from one's values collectively. Surely some higher level of agency is required to ensure that I value autonomously. But in answer it is sufficient to refer to the most basic account of responsibility that I proposed earlier. Responsibility for an action occurs when that action reflects on the agent *qua* agent. For the agent to be alienated from his scheme of values would be to suggest that there is some sense of the agent that might occur independently of the values that he has, that could somehow either generate and affirm, or reject and abandon, his complete set of values. It is very unclear exactly what that might refer to. As Watson notes, 'the important feature of one's evaluation system is that one cannot coherently dissociate oneself from it *in its entirety*'.[40] One can give up a set of values only in the light of some other set of values that one does accept. The scheme of values of the agent is so closely interwoven into the nature of the agent *qua* agent that there is no further possible account of agency that could be used to undermine responsibility.

This idea is sufficient to secure the basic idea that the hierarchy of value over desire can protect our theory of responsibility from the challenge of regress. When an agent performs an action, for that action to be free the following features must obtain. Firstly, it must have been guided by a desire. Secondly, that desire must not be alienated from the values of the agent. Thirdly, the value to which the desire is connected must not be alienated from the general system of values that the agent adopts.

A consequence of this account is that even if it is possible that there are cases where an individual is alienated from a particular value that he has, that does not reduce us to regress. To say that an individual is alienated from a particular value is to say that his other values do not have the necessary bearing on that individual value. For example, if the agent is brainwashed into valuing materialism, he may recoil from that sense of materialism. His valuing of materialism may be disconnected from the other values that he has, and yet we may not be able to expect that he does not value what he values. He may have reflected about that value as much as we may expect of him given the time available.

[39] This idea builds a little on a theory of incommensurable values that I have developed in 'Conflicts about Conflicts' (2002) 3–4, *Juridical Review* 183.

[40] 'Free Agency' 347.

If this is plausible, it merely involves recognition that values generally operate in a consistent set, from which one value may be alienated. In that case, it would be insufficient to secure the responsibility of an agent that he performed an action under a desire that was appropriately sensitive to some value that he holds. It would have to be shown further that his valuing was accepted by the agent in the light of his other values. What is not required for responsibility, because it is not conceivable, is that the complete set of values that the agent holds must be affirmed by some other level of agency. For it is not conceivable how this condition could possibly be fulfilled. An action reflects on an agent *qua* agent insofar as it reflects on the central features of agency. The idea of agency independent of a scheme of valuation lacks obvious content and for that reason it obviously cannot be central to a proper account of responsibility.

Furthermore, it is to be noted that, under this account, *any* individual value, if it is to ground responsibility, must be accepted in the light of other values that the agent adopts. Given that values tend to operate in the more or less systematic way that I suggest, each value that the agent holds must be at least accepted in the light of other values that the agent holds. If there are values that are resistant to such scrutiny and adjustment, in the way that I suggested may be true of brainwashing, they cannot ground attributions of responsibility. That is not to say that there is no hierarchy of values. There may be more particular values and more general values that an agent holds, and identification with the latter may be stronger than with the former. That fact does nothing to challenge the general account that I have provided. Even general values must be capable of scrutiny in the light of their more particular application in particular contexts, as well as with regard to other more general values.

1.3 Two Qualifications of the Refined Hierarchical Account

Responsibility, I have suggested, may be attributed when an agent acts on a desire that is accepted in the light of his values. However, the agent will not be responsible if the value under which the desire is accepted is not accepted in the general system of values that the agent has. This theory can accommodate the idea that alienation from one's values is possible. And it can do so without either positing a further level of agency that is independent of one's values, or succumbing to regress.

The possibility of regress is met by the idea that any value that the agent has, in order to be sufficiently connected to the agent to reflect on him *qua* agent, must be accepted in the light of other values that the agent has. That is not to say that all of the values of the agent must be coherent. If an agent merely fails at least to attempt to develop a coherent set of values, he is responsible for the values that he has, and any desires and actions that are accepted in the light of them. If an agent merely accepts a value which does not cohere with other values that he has, and acts on a desire that corresponds to that value, he is responsible for that action.

This leaves open the possibility that the agent may value something, but be alienated from that value. There may be values that the agent does not accept, and

consequently which he is not responsible for. Insofar as this is not the case, however, the agent's responsibility is secured by the relationship of any value that he holds to his system of values. The system of values that the agent has, on the other hand, is so intimately connected to the agent *qua* agent that the agent cannot be alienated from that system. There is no further level of agency in the light of which one can be alienated from one's system of values. That being the case, one cannot say that one's system of values does not reflect on one *qua* agent.

However, even this refined account will have to be qualified to make it plausible. The first qualification takes up the third objection to Frankfurt's model. I claimed that Frankfurt's model fails to appreciate the significance of time to responsibility. If an agent acts under a desire to *v* at *t*, and has a second-order volition with regard to *v*ing at *t*, the agent is responsible for *v*ing, Frankfurt claims. This is so, for Frankfurt, regardless of how he came to have such a desire and second-order volition. Frankfurt has been criticised for this.[41] A proper account of responsibility, it has been argued, ought to be historicised. I will take up this issue in more detail in Chapter 5, but it is worth introducing the issue here.

We can illuminate the significance of history by reflecting a little more on the basic claim that I have made about responsible agency. An agent is responsible for an action, I have suggested, if that action reflects on the agent *qua* agent. This invites us to investigate the nature of an agent *qua* agent, something which I will undertake in more detail in the next chapter. However, one thing that is immediately clear is that the identity of agents persists over time. For this reason whether a particular feature of the agent's psychology truly reflects his agency or not depends upon the history of the agent. A consequence of this is that whether a desire is reflective of agency cannot be understood simply by investigating the psychology of the agent at the time at which the action was performed. How the agent came to have that psychology will also be important in establishing responsibility.

Suppose that an agent has a compulsive desire at *t2*. She attempts to shake off that desire at *t2* but fails and she acts in accordance with it. If that desire was formed in accordance with her values at an earlier time, *t*, she may still be responsible for her actions at *t2*. Hence, the alcoholic who was too lazy to attend AA meetings at *t* might be responsible for her compulsive desire at *t2* even if, at *t2*, she attempted to control her desire for alcohol and failed. Such an agent has diachronically accepted her desire even if she has not synchronically accepted it.[42]

Now, it may be that if there is sufficient distance between the agent at *t* and the agent at *t2* in such cases, the agent is no longer responsible for the actions that she performs in the light of her compulsive desires at *t2*. Suppose that the agent at *t* forms a compulsive desire to *v*. At *t2* her set of values has fundamentally changed.

[41] See, for example, J M Fischer and M Ravizza *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: CUP, 1998) chs. 7 and 8 and, for a parallel account on the related topic of autonomy, A Mele *Autonomous Agents: From Self-Control to Autonomy* (Oxford: OUP, 1995) ch. 9.

[42] This is an adaptation of language proposed by Jeanette Kennett. In *Agency and Responsibility*, at 158, Kennett talks of diachronic and synchronic *control*, which, as I noted above, makes her account susceptible to Frankfurt-style objections.

However, between *t* and *t2* it was impossible to shake off the compulsive desire to *v*. In that case, her compulsive desire to *v* at *t2* may be sufficiently divorced from herself *qua* agent at *t2* such that she is no longer responsible for acting in accordance with that desire. Even voluntary addicts may gradually lose responsibility for actions performed due to their addiction over time if they become increasingly alienated from their desires. From this, we can see that our account of responsibility cannot be divorced from an investigation of the agent over time. In Chapters 5 and 12 I will show that a failure to recognise this historical feature of responsibility has led to an overly restrictive account of mental disorder defences.

A second qualification takes up an issue that I have already mentioned in relation to the distinction between free will and compulsion earlier in this chapter. There I suggested that there is a distinction to be drawn between a compulsive desire and a compulsive action. Suppose that D has a compulsive desire to *v*. He does not accept that desire, in the sense that I outlined above, and he has attempted to alter his desire to make it conform to his values. This attempt has failed. Furthermore, there has never been a time when D has accepted that desire. Consequently, he continues to desire something that he does not value. He then acts in accordance with that desire. Are we to see his action as compelled?

The answer, I think, depends upon further features of the situation. Suppose that there is a strong reason against D acting in accordance with his desire. Say, for example, in *v*ing D will also *r* and *r*ing is a serious wrong. However, D does not recognise that *r*ing is a serious wrong. Consequently he *v*s, thus also *r*ing. In this case, I think that it is right to say that D is responsible for his action. Although he has acted in accordance with a compelled desire, he has failed to recognise a strong reason against so acting. In this case, his action can be attributed to him *qua* agent. It is reflective of his insufficient recognition of the reasons against *r*ing. If, on the other hand, D had recognised *r* in the appropriate way and to the appropriate degree, in line with the requirements of normative reasons, and he had still *v*ed, then he would not have been responsible for *v*ing.

In this case, in contrast, D's *v*ing can in part be explained by his failure to recognise that *r*ing is a serious moral wrong, and that shows inadequate motivation on the part of D with respect to *r*. To return to the language of the first section of this chapter, it is appropriate to scrutinise D's motivating reasons in the light of normative reasons. And the distance between motivating and normative reasons generates D's responsibility for his action. This is so even though D was alienated from the reason that motivated the action. D's failure properly to recognise the significance of the normative reasons *against* acting in the way that he acted can ground his responsibility.[43]

In this chapter I have outlined the basic contours of the theory of responsibility that I will use to develop the principles of criminal responsibility throughout this

[43] We will consider some further instances of agents failing to recognise the significance of reasons against which they act in Chapter 8.

book. In order to achieve that I have used some of the ideas that have been generated by philosophical writing on the nature of free will and responsibility. Much of the philosophical discussion has not been treated with much care in accounts of criminal responsibility. Rather, theoretical writing on the general nature of criminal responsibility has tended to develop through the contrast between choice theories, capacity theories and character theories.

In the next chapter I hope to show how the theory of responsibility that I have generated here can throw some light on that theoretical literature. I will show that using this theory of responsibility allows us to accommodate the advantages of each of the popular theories of criminal law. It allows us to see the relevance of the temporal extension of the agent, which is an advantage of character theories, but it does this without requiring us to accept the implausible idea that an individual is not responsible for action that is performed 'out of character'. It results in a properly refined notion of respect for autonomy, without supporting the mistaken claim that an individual is only responsible where he could have chosen to have done otherwise than he did. And it allows us to refine the idea of what it means for an individual to have the relevant capacity to be regarded as criminally responsible.